MASS SPECTROMETRY-BASED PROTEOMIC ANALYSIS OF COMPLEX BIOLOGICAL SAMPLES

by

PENG ZHAO

(Under the Direction of Lance Wells)

ABSTRACT

Proteomics seeks to determine protein structure, modifications, localization, and protein-protein interactions in addition to protein expression levels. Although proteomics is often a large-scale endeavor on a global system, much of its attraction lies in the ability to focus its tools on selected populations of proteins in specific circumstances, contributing directly to address questions involving functions and mechanisms. With the advancement of analytical technology, mass spectrometry, which is capable of sequencing peptides/proteins and characterizing post-translational modifications (PTMs) in molecular details, has become the main power in proteomics. Furthermore, strategies developed to characterize individual proteins are now systematically applied to protein populations. To date, mass spectrometry has been used to map the complete primary structure of individual proteins, and is becoming a general method for the characterization of modified subproteomes in large-scale research.

Through the work outlined in this dissertation, we first explored the applications of mass spectrometry on human embryonic stem cells and discovered over 3000 proteins and 500 PTMs; later, a quantitative glycoproteomic study was performed on human pancreatic

ductal fluids and revealed several potential protein biomarkers for the early diagnosis of pancreatic cancer; finally, we developed and tested an HCD/ETD MS scheme for the site mapping of O-GlcNAc proteins, and further investigated the potential of its applicability to other types of protein PTM.

INDEX WORDS: mass spectrometry, proteomics, embryonic stem cell, phosphorylation, glycosylation, O-GlcNAc modification, glycoproteomics, pancreatic cancer, quantification, HCD/ETD, site mapping, post-translational modification

MASS SPECTROMETRY-BASED PROTEOMIC ANALYSIS OF COMPLEX BIOLOGICAL SAMPLES

by

PENG ZHAO

B.S. University of Science and Technology of China, 2005

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2011

MASS SPECTROMETRY-BASED PROTEOMIC ANALYSIS OF COMPLEX BIOLOGICAL SAMPLES

by

PENG ZHAO

Major Professor: Lance Wells
Committee: I. Jonathan Amster
Ron Orlando

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2011

DEDICATION

To my mother and father, who have always been reassuring and supportive; to my grandparents, who have instilled in me the merits of industriousness and perseverance; to my uncles, my aunts and my cousins, who have cared for me in more ways that I can ever imagine. Over the years, your encouragement has kept me motivated; your love has kept me comforted. You are what I call “family”, and there is nothing else in the world can compare to it.

And to my friends, from near and afar, who have been a constant inspiration. I treasure your friendship, and I’m grateful for your patience and understanding through the days of my frustration.

I know that don’t say this enough, but I’d like to start as of this moment: I love you all.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude and my appreciation to my mentor, Dr. Lance Wells. Thank you for taking a chance on me when you knew nothing about me, and giving me the opportunity to study and work in your laboratory and therefore exposing me to some brilliant minds and extraordinary science. You have been a magnificent teacher to me and I'm grateful for your constant guidance and support throughout the years of my graduate school.

I would also like to thank my committee members, Dr. Jon Amster and Dr. Ron Orlando, for offering your bright insight and expertise throughout this project.

There are also many people that have helped on providing me with their lab facility and their brain-storming discussions and many other unforgettable and greatly appreciated things. Among them, the past and present members of the Wells' laboratory, Enas Gad Elkarim, Jae-min Lim, Chin Fen Teo, Stephanie Hammond, Anu Rajesh Koppikar, Sandii Brimble, Edith Hayden, Meng Fang, Krithika Vaidyanathan, Sean Durning, Seongha Park, Ryan Stuart, Jeremy Praissman, and Robert Bridger, thank you all for your helpful suggestions and assistance. You are the ones that make this journey joyful, and I wish you all the best in your future endeavors.

Lastly but certainly not the least, I would like to thank my family and friends for lending me their unconditional love and support. My appreciation for you all is beyond my humble words. I love you all.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW 1

2 THE HUMAN EMBRYONIC STEM CELL PROTEOME REVEALED BY MULTIDIMENSIONAL FRACTIONATION FOLLOWED BY TANDEM MASS SPECTROMETRY 39

3 CANCER BIOMARKERS DISCOVERED IN PANCREATIC DUCTAL FLUID USING A GLYCOPROTEOMIC APPROACH 91

4 COMBING HIGH ENERGY COLLISION-INDUCED DISSOCIATION AND ELECTRON TRANSFER DISSOCIATION FOR PROTEIN O-GLCNAC MODIFICATION SITE ASSIGNMENT 123

5 CONCLUSIONS 169

APPENDICES

A RNA-GUIDED RNA CLEAVAGE BY A CRISPR RNA-CAS PROTEIN COMPLEX 172

B MOUSE AND ZEBRAFISH HOXA3 ORTHOLOGUES HAVE NONEQUIVALENT IN VIVO PROTEIN FUNCTION 213

C THE OUTER MEMBRANE PROTEOME OF BURKHOLDERIA PSEUDOMALLEI AND BURKHOLDERIA MALLEI FROM DIVERSE GROWTH CONDITIONS: INSIGHT INTO ABUNDANT PROTEINS ALWAYS PRESENT ON OR NEAR THE CELL SURFACE 241

LIST OF TABLES

Page

Table 2-1: Assignment of hES cell and secreted proteome and PTMs within hES cell proteome (1% FDR).....83

Table 2-2: Novel phosphorylation sites and corresponding proteins identified in hES cell proteome 84

Table 2-3: Implicated kinases predicted by Scansite with high-stringency filter.....85

Table 2-4: Novel O-GlcNAc modified proteins and corresponding sites.....86

Table 2-5: Proteins identified as being related to pluripotency of hESCs.....87

Table 2-6: Proteins that are modified by both phosphate and O-GlcNAc.....88

Table 2-7: List of identified secreted proteins of hESCs.....89

Table 2-S1: Comparison of specific protein sequence coverages between single LC-MS/MS and MudPIT experiments.....90

Table 3-1: Summary of the identified and quantified proteins in pancreatic ductal fluid samples.....119

Table 3-2: Variation of protein expression levels in pancreatitis, IPMN, and cancer samples compared to normal controls and variation of protein expression levels in IPMN and cancer compared to pancreatitis samples.....120

Table 3-3: Unique proteins observed in pancreatitis, IPMN and/or cancer on quantifiable and identifiable level.....122

Table 4-1: List of novel O-GlcNAc proteins identified in the enriched HEK293T sample that had not been observed in the previous experiment (Ref. 10) 166
Table 4-2: List of novel HexNAc modification sites identified in the enriched HEK293T sample. 167
Table 4-S1: Fragments of HexNAc and Hexose oxonium ions.....168
Table A-S1: Proteins identified by tandem mass spectrometry of native gel-purified RNA protein complexes, expanded from Figure A-1E.....211
Table B-1: Summary of the phenotypes, showing that Hoxa3mz allele functions virtually the same as Hoxa3zf allele.....240
Table C-1: 20 most abundant OM proteins of B. pseudomallei detected under all growth conditions.....265
Table C-2: 20 most abundant OM proteins of B. mallei detected under all growth conditions.....266

LIST OF FIGURES

Page

Figure 2-1: Experiment workflow 77

Figure 2-2: Neutral loss-dependent MS3 method.....78

Figure 2-3: Statistics across LC-MS/MS technically triplicate experiments.....79

Figure 2-4: Subcellular localization of proteins identified from hESCs.....80

Figure 2-5: Real time impedance-based assay of affects of interfering with CK2 or PKC signaling in hESC 81

Figure 2-6: Comparison between single LC-MS/MS and MudPIT experiments 82

Figure 3-1: Biological annotation of the identified proteins in pancreatic ductal fluid samples 113

Figure 3-2: Data filter process flow chart.....115

Figure 3-3: Variation of protein expression across pancreatic ductal fluid samples 116

Figure 3-4: Validation of proteomic data by immunoblotting.....118

Figure 4-1: Database search strategy for the enriched HEK293T sample.....145

Figure 4-2: Respective CID, ETD, and HCD spectra of standard O-GlcNAc modified peptides CKII and BPP 146

Figure 4-3: Corresponding HCD and ETD spectra of O-GlcNAc modified peptides identified in the enriched HEK293T sample 154

Figure 4-4: ETD spectrum of the co-eluted O-GlcNAc modified peptides.....158

Figure 4-S1: Multiple-engine database search strategy for the HEK293T sample 159

Figure 4-S2: HCD/ETD spectra of O-Mannose and O-GalNAc modified peptides 160
Figure A-1: Identification of a Ribonucleoprotein Complex Containing psiRNAs and Cas Proteins 200
Figure A-2: psiRNA Species in the RNP Contain a Common 5′ Sequence Element and Distinct 3′ Termini 203
Figure A-3: Specific Cleavage of Complementary Target RNAs 205
Figure A-4: Cleavage Occurs 14 Nucleotides from the 3′ Ends of the psiRNAs 206
Figure A-5: Target RNA Cleavage Requires Five Cmr Proteins and a Single psiRNA Species 208
Figure A-6: Model for the Function of psiRNA-Cmr Protein Complexes in Silencing Molecular Invaders 210
Figure B-1: Structure and expression of Hoxa3 alleles 235
Figure B-2: Zebrafish hoxa3a can substitute for mouse Hoxa3 in thyroid, ultimobranchial body, tracheal epithelium, and soft palate development.....236
Figure B-3: Cranial nerve, thymus, and parathyroid defects are not rescued by zfhoxa3a 237
Figure B-4: Novel pharyngeal skeleton morphologies in Hoxa3zf/zf and Hoxa3mz/mz mice, and skeletal phenotype of compound mutants with Hoxd3 238
Figure B-5: Hoxa3zf allele has null function in NCCs.....239
Figure C-1: SDS-PAGE and peptide mass fingerprint analysis of a purified OM preparation from B. pseudomallei.....264

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1. Protein and Protein Post-translational Modifications

Over the past century, there have been a number of revolutions in molecular biology, a prime examples is the exhaustive work done on whole-genome sequencing by the Human Genome Project [1-2]. Genomics provides researchers with sequence information of genes and therefore the ability to pursue comprehensive, global inquiries into the complex and intricate interworking of cells and organisms. Led by genomics, a variety of "-omics" disciplines have begun to emerge, each with unique contribution to both the structure and function of particular biomolecules and the mechanisms of specific cellular processes.

Genomics aims to determine the linear chromosomal sequence of model organisms, as well as sequence differences between individuals. Annotating the genome, including defining coding and regulatory sequences, is also part of genomics. In the last few decades, it has become widely recognized that the genome only represents the first layer of complexity in cellular systems. The next downstream "-omics" is transcriptomics, which studies the transcriptional regulation of genes by measuring their messenger levels via microarray hybridization technologies[3-6]. However, biological activity is not carried out by the static genome or the semi-dynamic transcriptome directly associated with genome expression. It is proteins, the third level of complexity, that act as the cellular building blocks and directly assert the potential function of genes

via enzymatic catalysis, molecular signaling, and physical interactions. Protein function depends on the precise amino acid sequence, the post-translational modification (especially regulatory ones such as phosphorylation), the three-dimensional structure, the protein concentration, the association/interaction with other proteins, and the extracellular environment, all of which cannot be determined by the genome sequence or the transcriptome level.

Accordingly, proteomics seeks to determine protein structure, modifications, localization, and protein-protein interactions in addition to protein expression levels[7]. Although proteomics is often a large-scale endeavor on a global system, much of its attraction lies in the ability to focus its tools on selected populations of proteins in specific circumstances, contributing directly to address questions involving functions and mechanisms.

The analysis of post-translationally modified subproteomes has become a significant aspect in mass spectrometry-based proteomics in the past few decades. Post-translational modifications (PTMs) are covalent processing events that change the properties of a protein by proteolytic cleavage or by addition of a modifying group to one or more amino acids. PTMs of a protein can modulate its activity and function by determining its localization, turnover, and interactions with other proteins. In signaling, for example, kinase cascades are turned on and off by the reversible addition and removal of phosphate groups[8]. Characterization of these modifications provides indispensable insight into the biological function of proteins and thus the cellular activities involving them, however, the analysis of modified subproteomes presents formidable challenges in proteomics.

With the advancement of analytical technology, mass spectrometry, which is capable of sequencing peptides/proteins and characterizing PTMs in molecular details, has become the main power in proteomics. Furthermore, strategies developed to characterize individual proteins are now systematically applied to protein populations. To date, mass spectrometry has been used to map the complete primary structure of individual proteins[9-11], and is becoming a general method for the characterization of modified subproteomes in large-scale research.

1.1 Phosphorylation

Current analytical protein methods show phosphorylation to be the most ubiquitous, evolutionary conserved PTM. The reversible and transient nature of protein phosphorylation allows signal transduction pathways to carry out diverse cellular functions. From bacteria to human, phosphorylation serves to modify protein function by altering protein stability, cellular location, substrate affinity, complex formation, and activity; thus allowing essential events such as cell cycle and growth to occur at precise time and locations. Important cellular activities such as DNA transcription and the regulation of cell cycle, and cellular processes such as neuronal migration and immune system recognition are all either initiated or controlled by protein phosphorylation events via the interplay of protein kinases and phosphatases. A hallmark of protein phosphorylation is its substoichiometric nature. Although at least one-third of all cellular proteins are estimated to be phosphorylated, their levels of phosphorylation vary widely and specific sites may be phosphorylated anywhere between less than 1% to greater than 90%.

Within the last few decades, there has been a tremendous effort in the design and advance of phosphoproteome detection techniques, such as ^{32}P labeling, Edman degradation, and mass spectrometry. Phosphorylation can be easily detected and visualized in 1D- and 2D-gels by ^{32}P labeling or by western blotting with phosphosite-specific antibodies, however, it is far more challenging to identify novel phosphoproteins and, in particular, to localize their phosphorylation sites. Moreover, in addition to simple detection of phosphorylation sites, quantitative studies of dynamic phosphorylation events are important to delineate and understand cell signaling pathways.

The versatility and speed of mass spectrometry makes it an ideal method for the analysis of protein phosphorylation. MS can, in principle, sequence each phosphopeptide, localize the phosphorylation sites, and acquire quantitative information. In the last few years, MS-based methods have become sufficiently sensitive and robust to be used routinely and in large-scale proteomic research. The application of MS in such field has resulted in the discovery of a large pool of phosphorylation sites from model organisms such as yeast, fly, mouse, and human[12-21].

Analysis of Phosphopeptides by Mass Spectrometry

The specific mass spectrometric properties of phosphopeptides create certain challenges during MS analysis. The fragmentation of phosphopeptides, specifically phosphoserine- and phosphothreonine-containing peptides, is compromised upon collision-induced dissociation (CID) in the tandem MS. During gas-phase CID, phosphate competes with the backbone as the preferred site of protonation and is subject to nucleophilic attack from a neighboring amide carbonyl group upon collision. As a result, the phosphate is often cleaved leaving as a phosphoric acid and an energetically

favorable α, β-unsaturated carbonyl group is subsequently formed. This loss of phosphoric acid is often so predominant that additional fragmentation of sequence revealing peptide backbones becomes a secondary process and is thus greatly suppressed[22]. Also, there is possible gas-phase rearrangement during the fragmentation, which may cause incorrect phosphosite localization[23-24].

An alternative fragmentation method is electron-based dissociation, such as electron capture and electron transfer dissociation (ECD and ETD, respectively) [25-27]. Electron-based methods induce cleavage of the backbone N-Cα bond via free radical ion/ion chemistry, generating c- and z-ions for peptide sequencing. Many PTMs that are labile upon CID, such as phosphorylation, glycosylation, sulfonation, etc., are preserved during ECD and ETD [28-35]. CID and ETD fragmentation techniques have been proven complementary in analyzing proteins and phosphoproteins from complex biological samples [33, 36].

1.2 Glycosylation

Glycosylation, characterized by the attachment of oligosaccharides to proteins through amide nitrogen (N-linked) or hydroxyl oxygen (O-linked), is one of the most important and common forms of protein post-translational modification. Protein glycosylation is involved in many physiological functions and biological pathways, and altered glycosylation has been associated with a variety of human pathology, including cancer, inflammatory and degenerative diseases. Glycoproteins are becoming important targets for the development of biomarkers for disease diagnosis, prognosis and therapeutic response to drugs. The emerging technology of glycoproteomics, which focuses on glycoproteome analysis, is increasingly becoming an important tool for

biomarker discovery. An in-depth, comprehensive identification of aberrant glycoproteins, and further, quantitative detection of specific glycosylation abnormalities in a complex environment require a concerted approach drawing from a variety of techniques.

With recent advances in analytical and computational technologies, glycoproteomics, the global analysis of glycoproteins, is rapidly emerging as a sub-field of proteomics with high biological and clinical relevance. Glycoproteomics integrates glycoprotein enrichment and proteomics technologies to support the systematic identification and quantification of glycoproteins in a complex sample. The recent development of these techniques has stimulated great interest in applying the technology in clinical translational studies, in particular, protein biomarker research.

Analysis of Glycoproteome by Mass Spectrometry

Mass spectrometry, due to its high sensitivity and selectivity, has been one of the most versatile and powerful tools in glycoprotein analysis to identify glycoproteins, assign glycosylation sites and elucidate oligosaccharide structures [37-39]. The utility of a top-down approach (intact protein based proteomics analysis) [40] for glycoprotein characterization in a complex sample is still technically challenging with current technology. At present, the most widely used glycoproteomics methods are based on characterizing glycopeptides derived from glycoproteins, analyzing either deglycosylated glycopeptides or intact glycopeptides with glycan attachment.

The direct analysis of intact glycopeptides with carbohydrate attachments is complicated by the mixed information obtained from the fragment ion spectra, which may include fragment ions from the peptide backbone, the carbohydrate group and the

combinations of both. While it is technically challenging to comprehensively analyze intact glycopeptides on a global scale for complex biological samples, complementary information regarding peptide backbone and glycan structure can likely be obtained in a single measurement. Early work using CID has identified a few key features that are characteristics of the fragmentation of glycopeptides, providing the basis for intact glycopeptide identification [41-45]. The analysis of intact glycopeptides has been carried out using a variety of different instruments, including electrospray ionization (ESI) based ion trap (IT) [46-50], quadrupole ion trap (QIT) [51-53], Fourier transform ion cyclotron resonance (FT-ICR) [54-57], ion trap time-of-flight (IT-TOF) [58-59] and quadrupole time-of-flight (Q-TOF) [60-65]; matrix-assisted laser desorption/ionization (MALDI) based Q-TOF [66-68], quadrupole ion trap time-of-flight (QIT-TOF) [52, 69-70] and tandem time-of-flight (TOF-TOF) [49-50, 69, 71-73] mass spectrometers.

In general, the CID generated MS/MS spectrum of a glycopeptides is dominated by B- and Y-type glycosidic cleavage ions (carbohydrate fragments) [74], and b- and y-type peptide fragments from the peptide backbone. However, the MS/MS fragmentation data obtained from different instruments can have pronounced difference in providing structure information on glycan and peptide backbone depending on the experimental setting and instrumentation used for the analysis, including ionization methods, fragmentation techniques and mass analyzers. Low energy CID with ESI based ion trap, FT-ICR and Q-TOF instrument predominantly generates fragments of glycosidic bonds. The increase of collision energy using FT-ICR and Q-TOF instrument results in more efficient fragmentation of b- and y- ions from peptide backbone. MALDI ionization generates predominantly singly charged precursor ions, which are more stable and

usually fragmented using higher energies via CID or post-source decay (PSD), generating fragments from both the peptide backbone and the glycan [66-68, 73, 75-78]. While Q-TOF instruments have been widely used for intact glycopeptide characterization, one unique feature of the ion trap instrument is that it allows repeated ion isolation and fragmentation cycles, which can provide a wealth of complementary information to interpret the structure of glycan moiety and peptide backbone [39, 52, 79]. Recently, fragmentation techniques using different mechanisms from CID have been introduced and applied for glycopeptide analysis, including infrared multi photon dissociation (IRMPD) [80-83], ECD [80-88] and ETD [51, 89-91]. The application of IRMPD and ECD is largely performed with FT-ICR instruments. Complementary to CID fragmentation, ECD and ETD tend to cleave the peptide backbone without the loss of glycan moiety, providing specific information on localizing the sites of glycosylation.

While great efforts have been made to apply a variety of mass spectrometry techniques to study both N-linked [39, 52-53, 80-82, 85, 92-97] and O-linked [58, 84-85, 87-88, 98-107] glycopeptides, the interpretation of the fragment spectrum of an intact glycopeptide still requires intensive manual assignment and evaluation. A recent study has demonstrated the feasibility to develop an automated workflow for analyzing intact glycopeptides in mixtures [108]. In general, however, a high throughput, large scale profiling of intact glycopeptides in complex samples still remains a challenge with current technology.

The analysis of deglycosylated peptides requires the removal of glycan attachments from glycopeptides. For N-linked glycopeptides, the glycosidic bond can be specifically cleaved using the enzyme PNGase F or PNGaseA, providing deglycosylated peptides which can then be analyzed directly via shotgun proteomics. The PNGase F/A-

catalyzed deglycosylation results in the conversion of asparagine to aspartic acid in the glycopeptide sequence and introduces a mass difference of 0.9840 Da, which can be used to precisely map the N-linked glycosylation sites using high resolution mass spectrometers. Stable isotope labeling of glycepeptides in $H_2{}^{18}O$ has also been used in combination to the enzymatic cleavage of glycans to enhance the precise identification of N-glycosylation sites [109-111]. The removal of O-linked glycans is less straightforward, and mostly relies on chemical methods, such as trifluoromethansulfonic acid [112], hydrazinolysis [113], β-elimination [114], and periodate oxidation [115-116]. The application of these methods suffers from a variety of limitations, such as low specificity for O-linked glycosylation, degradation of the peptide backbone, and modifications of the amino acid residues, all of which can complicate or compromise O-linked glycosylation analysis in complex samples. Most of the large-scale glycoproteomics studies using the deglycosylation approach have been focused on N-glycoproteins, which are prevalent in blood and a rich source for biomarker discovery. O-glycosylation lacks a common core, consensus sequence and universal enzyme that can specifically remove the glycans from the peptide backbone, thus, is more challenging to analyze for large scale profiling.

2. Quantitative proteomics

Another important application of proteomics focuses on the quantitative analysis of proteins.

Until recently, proteomics was largely a qualitative discipline. Typical proteomic experiments resulted in lists of proteins identified at a given state as a snapshot of certain biological system without any further information about abundance, distributions, or stoichiometry. On contrary, the population of proteins is seldom static but rather

dynamic, regulating biological activities via numerous pathways. In order to allow interpretation of biological systems in a real-time fashion, quantitative analyses are necessary.

The applications of quantitative proteomics are equally diverse, ranging from the quantitative evolution of a protein over time or in disease progression, to the direct comparison of samples, such as normal and diseased tissues. Each of these applications comes with its own characteristics and expectations, and a broad set of different protocols has been developed over the last few years in order to address their varied goals.

Both relative and absolute quantification can be achieved in proteomic analysis. Specifically, the methods in mass spectrometry-based quantitative proteomics can be divided in two main families: stable isotope labeling and label-free methods. As implied in the name, the first family employs differential stable isotope labeling to introduce a specific mass difference that can be recognized by mass spectrometers providing the basis for quantification. These mass differences can be introduced into proteins or peptides (i) metabolically, (ii) chemically, (iii) enzymatically, or (iv) via the incorporation of certain synthetic peptide standards. In contrast, label-free approaches aim to correlate the mass spectrometric signal of either intact peptides or characteristic fragments of peptides, or the number of peptide sequencing events with the relative or absolute protein quantity.

2.1 Stable isotope labeling methods

Labeling of peptides by stable isotopes allows the quantification of peptides to take place directly in an MS spectrum, since the isotope label is perceived by the mass spectrometer as a shift in the mass-to-charge ratio (m/z). Basically, isotope labeling

methods are dedicated to relative quantification (labeled versus unlabeled or comparisons between different labels). However, with the appropriate experimental design, the same method can readily be used for absolute quantification, which is typically accomplished by introducing a labeled synthetic peptide at known concentration and comparing its intensity to that of the unlabeled peptide.

Using isotopic labeling approaches, the quantitative information can be obtained from either MS1 (parent ion) or MS2 (daughter ion) spectra, in the latter of which the reporter ions are typically involved and can only be released upon fragmentation. Furthermore, a multiplex method also exists, which builds upon the same labeling strategy used in the MS1 methods but quantifies based on the information in the MS2 spectra.

2.1.1 MS1 methods

These quantification methods are based on the differentiation of samples at the MS1 level. In MS1 methods, stable isotopes with different masses are introduced to label samples so that a common peptide found in more than one sample will yield multiple isotopic envelopes in the MS1 spectra, allowing the differential quantification of the samples.

2.1.1.1 Stable-isotope labeling with amino acids in cell culture (SILAC) (in vivo labeling)

SILAC is a metabolic labeling strategy that uses stable isotope-labeled amino acids in growth medium to encode cellular proteomes for quantitative analysis [117]. The technique depends on cellular protein synthesis to incorporate stable isotope-containing amino acids into whole proteome. Basically, two populations of cells are grown in two

separate medium formulations, the light medium containing the amino acid with the natural isotope abundance and the heavy medium containing the heavy isotope labeled amino acid. After a few replication cycles, more than 97% or even full incorporation of the heavy amino acid can be reached in the proteomes [118]. The proteins or peptides from unlabeled and labeled samples are subsequently mixed together and analyzed by MS, which are distinguished on the spectra by a residue-specific mass difference corresponding to the number of stable isotope labels as well as the number of labeled amino-acid residues in the analyte. The signal intensities from light and heavy samples provide a quantitative comparison of their relative abundances in the mixture.

SILAC is a simple, robust, and powerful approach in mass spectrometry (MS)-based quantitative proteomics. The early combination of samples after labeling in cell culture means that they will be subjected to identical downstream workflow throughout sample processing and analysis, thus minimizing technical variability [118].

Like other in vivo labeling methods, there is the possibility that by altering the metabolic state of the living environment of cells will have effects on the proteome.

2.1.1.2 ^{18}O Labeling

The ^{18}O labeling method [119] relies on the differentiation of two samples using different stable isotopes of oxygen. The heavy ^{18}O isotope is introduced into the C-terminus of peptides by performing the digestion in $H_2{}^{18}O$. The proteolytic reaction involved in ^{18}O labeling can be performed in a variety of enzymes, such as trypsin, LysC, GluC, etc.

The reactions involved are:

$$RC^{16}ONHR' + H_2O \rightarrow RC^{16}O^{18}O^- + H_3NR' \quad ①$$

$$RC^{16}O^{18}O^{-} + H_2^{18}O \rightarrow RC^{18}O^{18}O^{-} + H_2^{16}O \quad ②$$

There are many advantages of using ^{18}O labeling method: the reaction is simple and uniform, every peptide will be labeled equally and the resulting 4 Da mass shift allows labeled peptides to be distinguished from unlabeled ones. Furthermore, the method is universal, as it can be applied to any protein sample.

One common problem with this method is the occurrence of background exchange, which results in the reintroduction of ^{16}O into previously fully labeled peptide C-termini. Background exchange is mainly caused by the slow accumulation of OH in the digestion buffer due to labeling (reaction ②). The result of the background exchange is an overall reduction in efficiency leading to the rise of a population of peptides labeled with only one ^{18}O and a very small minority of completely unlabeled peptides. The presence of the half-labeled peptides actually makes the distinction between samples much more complicated, as the isotope envelopes of the three forms start to overlap, whereas the unlabeled peptides would be indistinguishable from the peptides in the unlabeled sample. This background exchange problem can be addressed by using a protocol that stops this process in its tracks [120].

2.1.1.3 Isotope-coded affinity tag (ICAT)

ICAT was initially developed by Gygi et al. in 1999 [121], which employs tags that consist of three elements: an affinity tag (biotin), which is used to isolate ICAT-labeled peptides; a linker that can incorporate stable isotopes, which exists in heavy (contains eight deuteriums) and light (contains no deuteriums) forms; and a reactive group with specificity toward thiol groups (cysteines). The reactive group ensures the labeling of reduced cysteines, and the affinity tag is used to isolate ICAT-labeled peptides from a

proteolytic digest. The linker that connects the two functional groups is loaded with eight deuteriums or hydrogens, which establishes an 8 Da mass shift in the MS1 spectrum between the heavy- and light-labeled peptides.

The main drawbacks of ICAT are adverse side reactions and its inability to label peptides that do not contain cysteine. On the other hand, it is a well established and widely used method that does not require much post-processing.

2.1.1.4 Isotope-coded protein labels (ICPL)

The isotope-coded protein labels (ICPL) method relies on the same principle as ICAT of labeling the peptides with reagents of different masses. In 2004 Schmidt et al. employed the method to quantify proteins in a dynamic range from 0.25 to 30 with a coefficient of variation under 10% [122]. Compared to ICAT, this method has the advantage of labeling all peptides as the reactive group targets free amines, whereas ICAT presents a mechanism for peptide selection through the affinity tag that ICPL does not support.

Usually ICPL reagents are used to compare two proteomes. However, taken together the four ICPL reagents allow multiplexing of up to four samples. The samples will then be distinguished by mass differences of 4, 6 and 10 Da compared with the ICPL-d0 carrying peptides.

2.1.1.5 Absolute quantification (AQUA)

The AQUA method [123] differs from the previous methods in that quantification is performed on a known/standard peptide. AQUA labeling is therefore used to monitor the quantitative behavior of a specific (set of) protein(s). The method starts by synthesizing peptides, with/without covalent modifications (e.g., phosphorylation, methylation, acetylation, etc.), that originate from the target proteins using stable isotopes (e.g. ^{13}C,

^{15}N, etc.). The synthetic peptides are then introduced into the experimental sample at known concentrations as an internal standard, and used to measure the absolute levels of proteins and/or post-translationally modified proteins after proteolysis via a selected reaction monitoring (SRM) analysis. At MS1 stage, the synthesized and labeled peptide can be distinguished from the native one with a mass shift of 6-10 Da, and its intensity in relation to that of the native peptide can be used to quantify the targeted protein.

This targeted approach has made it possible to quantitatively investigate proteins/peptides at low abundances, where the signal can easily be lost in the noise or lies outside of the dynamic range of the instrument. The overall strategy is particularly useful for SRM experiments, which implicitly targets known peptides and a set of their fragments.

2.1.2 MS2 methods utilizing reporter ions

The reporter ion methods were introduced in quantitative proteomics a few years ago. The basic concept is to label the peptides in each sample with a different label molecule after protein digestion and to quantify them based on the intensity of their specific fragment ions generated upon MS2. Typically, each of the labels consists of an identical reaction group that can bind to a peptide, a variable spacer group and a variable reporter group which balance each other out so that different labels remain isobaric. As a result, a common peptide occurring in different samples and thus carrying different labels in the final mixture will still present as a single peptide isotopic envelope in an MS1 spectrum, possibly at an increased m/z corresponding to the constant mass increase contributed by the labels. The labels are designed to fragment efficiently upon MS2 so

that the reporter moiety is released, and the quantification is achieved by calculating the signal intensity at each of the unique/expected masses carried by the reporter groups.

Quantification at the MS2 level is particularly beneficial in the following ways: for the analysis of complex samples, the initial precursor selection inherent in MS2 analyses implicitly ignores a substantial part of the background signal; also, identification is not required in order to quantify the signal of a given precursor across samples, since it is sufficient to collect the intensity of various reporter ions which can be found at known masses; finally, information on relative peptide quantification becomes available, which can be subsequently compiled into protein ratios after protein inference.

A shortcoming comes from the nature of the bonds between the reporter and balance group: most of the bonds are designed to be cleaved upon collision-based fragmentation such as CID or HCD, which makes it incompatible with electron-related dissociations such as ECD or ETD. Also, larger labels often have lower reactivity.

2.1.2.1 Isobaric tag for relative and absolute quantification (iTRAQ)

Over the past few years, iTRAQ has become one of the most popular quantification methods in proteomics. Introduced by Ross et al. in 2004 [124] and distributed by Applied Biosystems, this reporter method allows the relative and absolute quantification at the MS2 level of up to eight samples.

iTRAQ labeling reagents consist of three parts that will be separated during fragmentation: The peptide reactive group that reacts with primary amines, the reporter group (m/z can be one of 113, 114, 115, 116, 117, 118, 119 or 121) and the balance group that will ensure a constant mass for the intact label molecule.

This method is highly reproducible, and the isobaric nature of the labels allows simultaneous comparison of multiple samples in LC-MS/MS experiments with no increase in chromatographic or MS complexity. The labeling reaction is global for that any peptide with a free amine group can be labeled and measured, and the peptide coverage is significantly increased compared to ICAT method [124].

2.1.2.2 Tandem mass tag (TMT)

The TMT molecule has a structure made of four main modules: a peptide reactive group, a spacer (normalizer), a reporter, and a specific linker group to ensure good reporter cleavage upon fragmentation [125]. The reactive group can be modified to allow different biological nucleophiles to be labeled, also to accommodate an affinity ligand, such as biotin [125].

TMTs are designed to give rise to fragments with specific m/z ratio regardless of charge states, and allow simultaneous determination of both the identity and relative abundances of peptide pairs using CID-based MS/MS analysis. The simplest form available is TMT0, which is dedicated to the testing and optimization of the method at low cost, containing no differential labeling at all. Furthermore, more than two tags can be generated (up to TMT6) to allow for multiple samples.

2.1.3 Multiplex method

The multiplex method also uses an isotope to distinguish two different samples by heavy- and light-labeled peptides [126]. Compared to typical MS1 methods, which uses a small precursor window to isolate a single peptide and obtains MS2 data for protein identification whereas quantification is based directly or indirectly on MS1 data, the multiplex method does not record separate MS2 spectra for the heavy and light peptides

but a single MS2 spectrum for the peptide pair using a larger precursor window, and both protein identification and quantification are based on the multiplexed MS2 data.

The wide precursor selection window employed in this method does not significantly decrease the quality of the MS2 spectra; neither does the position of the ion in the window as demonstrated in [127] where a window of m/z 10 is introduced. This m/z 10 window gave more identifications than an m/z 2.5 window, and from m/z 2.5 to 10 the intensity increased until it reached a maximum. Overall, this method offers several advantages over conventional MS1 strategies, including increased sensitivity for protein identification and more accurate quantification with an extended dynamic range. In addition, the quantification process is fast, fully automated, and independent of instrument data type. However, it should be noted that increasing the window could decrease the selectivity of the method, which is a sensitive issue when analyzing complex samples.

2.2 Label-free quantification

Label-free quantification aims to provide quantitative information without introducing any form of labeling. The principle is to find relevant indicators of (relative) protein abundance directly in the mass spectrometer output. Different methods have been developed based on different types of MS signal.

Most label-free methods compare two or more samples based on the ion intensities of identical peptides or based on the number of acquired spectra for each protein. Ideally, samples for label-free comparisons are run consecutively on the same LC-MS/MS setup to avoid variations in ion intensities due to differences in the system

setup (column properties, temperatures), and thereby allow precise reproduction of retention times.

Label-free approaches present several advantages over labeled strategies: they are inexpensive and requires less sample preparation; they can be applied to any biological material; and the proteome coverage of quantified proteins is often high because basically every protein that is identified by one or more peptide/spectra can be quantified; the complexity of the sample will not be increased since mixing of different proteomes is not required. Therefore, label-free methods usually have a high analytical depth and large dynamic range, making it especially advantageous when large-scale/global-wise proteomic changes between biological samples are expected.

However, label-free methods are more demanding on the instrumentation, as the quality of the quantification relies heavily on the quality of the data and high precision and accuracy therefore become essential, especially with complex mixtures which require both high-resolution mass spectrometers as well as state of the art data processing tools.

2.2.1 Spectral counting

The spectral counting method uses the number of peptide-identifying spectra assigned to each protein as a quantitative measure [128]. This method assumes a linear relationship between the level of sampling observed for a protein and its relative abundance. Such linearity was observed to hold over two orders of magnitude [128]. The outcome of spectral counting depends on the settings of data-dependent acquisition on the mass spectrometer. In particular, the linear range for quantification and the number of proteins to be quantified are influenced by different settings for dynamic exclusion [129]; the optimal settings depend on sample complexity.

The most significant disadvantage of spectral counting is that it behaves very poorly with proteins of low abundance and few spectra. The accuracy of this method, especially for low abundance proteins, suffers from the fact that each spectrum is scored with the value "1" independently of its ion intensities. To overcome this problem, an approach has been suggested that uses the average of the total ion count from all fragment spectra that identify a protein as a quantitative measure. Thereby, the linear dynamic range for quantification can be vastly exceeded [130].

2.2.2 Protein abundance index (PAI, emPAI, and APEX)

The empirical relationship between the number of spectra or peptides identified for a given protein and overall protein abundance in the sample has been used as a basis to calculate the absolute concentration of each protein within the sample. The PAI of a protein is defined as the ratio between the sequenced peptides of a protein and the total number of tryptic peptides predicted from the protein sequence, which allows an approximate quantification of proteins based on the results of peptide identification [131]. Exponentially modified protein abundance index (emPAI) has been subsequently introduced as a more reliable and robust quantification indicator [132]. A very similar approach was taken to calculate the absolute protein expression (APEX) index [133]. The estimated protein concentrations calculated by emPAI indices have correlated very well with the protein concentrations calculated from enzymatic activities [134]. The indices emPAI and APEX are thus derived measures of absolute protein abundance in a given sample based on the analytical features in mass spectrometric analysis.

A benefit of this method is that it only requires the identification results, and the data processing is correspondingly simple. Its simplicity has led to ready adoption, even

in commercial tools; for instance, emPAI will be calculated by MASCOT for every database search containing more than 100 spectra.

2.2.3 TIC

The TIC, sum of the ion currents, can be measured or calculated in order to target specific ions [135]. While the length of proteins might bias spectral counting quantification, TIC-based quantification has been shown to be more accurate and to expand the dynamic range. The work done by Asara et al. [130], when the concentration ratio is under 20:1, TIC and spectral counting gave out nearly identical measurement; however, the results differed dramatically at high ratios and the result from TIC suggested an increase of the dynamic range.

2.2.4 Replicate

Spectral counting, emPAI and TIC were primarily used on low-resolution mass spectrometers, where they provide a relatively simple and fast quantification. However, their performance varies and can often prove insufficient. A higher accuracy method was found in the analysis of extracted ion chromatograms, which used the total intensity of the corresponding precursor measured at the MS1 level to represent the concentration of a peptide [136].

This method can be applied to known peptides, which will be targeted in the MS1 spectrum but also without prior peptide identification. Precursors are detected in the MS1 spectrum by a feature finder algorithm (parent mass list). The areas of the peaks matched to the pattern will then be used to calculate the concentration of the peptide.

With high-resolution mass spectrometers and reproducible peptide separation, peptides from different samples that were each recorded in a separate run (technical replicate) can be associated and quantitatively compared with each other [137].

This label-free quantification method offers a wide dynamic range as well as a high precision and accuracy. Furthermore, it can be used to compare complex samples for biomarker discovery, and is compatible with SRM experiments. However, this method does rely very strongly on complex data processing algorithms.

2.2.5 Average intensity method

The average intensity method is dedicated to absolute quantification. It relies on the intensity of the three most intense tryptic peptides for a protein. Proteins can be quantified across two orders of magnitude with a CV (coefficient of variation) of less than 30% [138].

A single point calibration is set for the mass spectrometer in this method to act as a reference for the absolute quantification.

2.3 Selected/Multiple reaction monitoring (SRM/MRM)

SRM or MRM is increasingly applied to quantitative proteomics because of its selectivity (two levels of mass selection), its sensitivity (non-scanning mode), and its wide dynamic range.

A triple quadrupole (QQQ) mass spectrometer is typically used for these analyses, although ion trap instruments can be used as well. In a QQQ-based SRM experiment, the first and the third quadrupoles act as filters to specifically select predefined m/z values corresponding to the peptide ion and a specific fragment ion of the peptide, whereas the second quadrupole serves as collision cell. Several such transitions (precursor/fragment

ion pairs) are monitored over time, yielding a set of chromatographic traces with the retention time and signal intensity for a specific transition as coordinates. The two levels of mass selection with narrow mass windows result in a high selectivity, as co-eluting background ions are filtered out very effectively. The non-scanning nature (no full MS is recorded) of the SRM operation leads to an increased sensitivity by one or two orders of magnitude compared with conventional full MS techniques. In addition, it results in a linear response over a wide dynamic range up to five orders of magnitude. This enables the detection of low-abundance proteins in highly complex mixtures, which is crucial for systematic quantitative studies [139].

In QQQ-based SRM analyses, the width of the mass window configured in the first quadrupole as well as the collision energy has an important impact on the quality of the quantification. A larger window will provide a higher intensity but also a lower signal-to-noise ratio. Reducing the mass window will decrease the overall intensity but may offer better selectivity on the transition.

The numerous methods developed for the quantification in proteomics provide efficient protocols that can be applied to a variety of problems.

Through the following three chapters, the work of investigating and developing MS-based methodologies in protein and PTM identification as well as quantification will be presented, including a large-scale proteomic study of human embryonic stem cells, a quantitative analysis of glycoproteins as biomarker candidates in pancreatic ductal fluids, and a targeted MS approach for interrogating the O-GlcNAc modified proteins.

REFERENCES

1. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J., Initial sequencing and analysis of the human genome. *Nature* **2001,** 409, (6822), 860-921.

2. Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.;

Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigo, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X., The sequence of the human genome. *Science* **2001,** 291, (5507), 1304-51.

3. Bowtell, D. D., Options available--from start to finish--for obtaining expression data by microarray. *Nat Genet* **1999,** 21, (1 Suppl), 25-32.

4. Cheung, V. G.; Morley, M.; Aguilar, F.; Massimi, A.; Kucherlapati, R.; Childs, G., Making and reading microarrays. *Nat Genet* **1999,** 21, (1 Suppl), 15-9.

5. Duggan, D. J.; Bittner, M.; Chen, Y.; Meltzer, P.; Trent, J. M., Expression profiling using cDNA microarrays. *Nat Genet* **1999,** 21, (1 Suppl), 10-4.

6. Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995,** 270, (5235), 467-70.

7. Yarmush, M. L.; Jayaraman, A., Advances in proteomic technologies. *Annu Rev Biomed Eng* **2002,** 4, 349-73.

8. Cohen, P., The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem Sci* **2000,** 25, (12), 596-601.

9. Roepstorff, P., Mass spectrometry in protein studies from genome to function. *Curr Opin Biotechnol* **1997,** 8, (1), 6-13.

10. Ling, V.; Guzzetta, A. W.; Canova-Davis, E.; Stults, J. T.; Hancock, W. S.; Covey, T. R.; Shushan, B. I., Characterization of the tryptic map of recombinant DNA derived tissue plasminogen activator by high-performance liquid chromatography-electrospray ionization mass spectrometry. *Anal Chem* **1991,** 63, (24), 2909-15.

11. Carr, S. A.; Hemling, M. E.; Folena-Wasserman, G.; Sweet, R. W.; Anumula, K.; Barr, J. R.; Huddleston, M. J.; Taylor, P., Protein and carbohydrate structural analysis of a recombinant soluble CD4 receptor by mass spectrometry. *J Biol Chem* **1989,** 264, (35), 21286-95.

12. Zhai, B.; Villen, J.; Beausoleil, S. A.; Mintseris, J.; Gygi, S. P., Phosphoproteome analysis of Drosophila melanogaster embryos. *J Proteome Res* **2008,** 7, (4), 1675-82.

13. Mann, K.; Olsen, J. V.; Macek, B.; Gnad, F.; Mann, M., Phosphoproteins of the chicken eggshell calcified layer. *Proteomics* **2007,** 7, (1), 106-15.

14. Macek, B.; Mijakovic, I.; Olsen, J. V.; Gnad, F.; Kumar, C.; Jensen, P. R.; Mann, M., The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. *Mol Cell Proteomics* **2007,** 6, (4), 697-707.

15. Macek, B.; Gnad, F.; Soufi, B.; Kumar, C.; Olsen, J. V.; Mijakovic, I.; Mann, M., Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics* **2008,** 7, (2), 299-307.

16. Villen, J.; Beausoleil, S. A.; Gerber, S. A.; Gygi, S. P., Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A* **2007,** 104, (5), 1488-93.

17. Li, X.; Gerber, S. A.; Rudner, A. D.; Beausoleil, S. A.; Haas, W.; Villen, J.; Elias, J. E.; Gygi, S. P., Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J Proteome Res* **2007,** 6, (3), 1190-7.

18. Dephoure, N.; Zhou, C.; Villen, J.; Beausoleil, S. A.; Bakalarski, C. E.; Elledge, S. J.; Gygi, S. P., A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A* **2008,** 105, (31), 10762-7.

19. Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P., Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* **2004,** 101, (33), 12130-5.

20. Ballif, B. A.; Villen, J.; Beausoleil, S. A.; Schwartz, D.; Gygi, S. P., Phosphoproteomic analysis of the developing mouse brain. *Mol Cell Proteomics* **2004,** 3, (11), 1093-101.

21. Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M., Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006,** 127, (3), 635-48.

22. Carr, S. A.; Huddleston, M. J.; Annan, R. S., Selective detection and sequencing of phosphopeptides at the femtomole level by mass spectrometry. *Anal Biochem* **1996,** 239, (2), 180-92.

23. Aguiar, M.; Haas, W.; Beausoleil, S. A.; Rush, J.; Gygi, S. P., Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets. *J Proteome Res* **2010,** 9, (6), 3103-7.

24. Palumbo, A. M.; Reid, G. E., Evaluation of Gas-Phase Rearrangement and Competing Fragmentation Reactions on Protein Phosphorylation Site Assignment Using Collision Induced Dissociation-MS/MS and MS(3). *Anal Chem* **2008**.

25. Coon, J. J.; Syka, J. E. P.; Schwartz, J. C.; Shabanowitz, J.; Hunt, D. F., Anion dependence in the partitioning between proton and electron transfer in ion/ion reactions. *International Journal of Mass Spectrometry* **2004,** 236, (1-3), 33-42.

26. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **2004,** 101, (26), 9528-33.

27. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W., Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* **1998,** 120, (13), 3265-3266.

28. McAlister, G. C.; Berggren, W. T.; Griep-Raming, J.; Horning, S.; Makarov, A.; Phanstiel, D.; Stafford, G.; Swaney, D. L.; Syka, J. E.; Zabrouskov, V.; Coon, J. J., A

proteomics grade electron transfer dissociation-enabled hybrid linear ion trap-orbitrap mass spectrometer. *J Proteome Res* **2008,** 7, (8), 3127-36.

29. Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E.; Coon, J. J., Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal Chem* **2007,** 79, (2), 477-85.

30. Shi, S. D.; Hemling, M. E.; Carr, S. A.; Horn, D. M.; Lindh, I.; McLafferty, F. W., Phosphopeptide/phosphoprotein mapping by electron capture dissociation mass spectrometry. *Anal Chem* **2001,** 73, (1), 19-22.

31. Khidekel, N.; Ficarro, S. B.; Clark, P. M.; Bryan, M. C.; Swaney, D. L.; Rexach, J. E.; Sun, Y. E.; Coon, J. J.; Peters, E. C.; Hsieh-Wilson, L. C., Probing the dynamics of O-GlcNAc glycosylation in the brain using quantitative proteomics. *Nat Chem Biol* **2007,** 3, (6), 339-48.

32. Kelleher, N. L.; Zubarev, R. A.; Bush, K.; Furie, B.; Furie, B. C.; McLafferty, F. W.; Walsh, C. T., Localization of labile posttranslational modifications by electron capture dissociation: the case of gamma-carboxyglutamic acid. *Anal Chem* **1999,** 71, (19), 4250-3.

33. Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J., Performance characteristics of electron transfer dissociation mass spectrometry. *Mol Cell Proteomics* **2007,** 6, (11), 1942-51.

34. Molina, H.; Horn, D. M.; Tang, N.; Mathivanan, S.; Pandey, A., Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A* **2007,** 104, (7), 2199-204.

35. Chi, A.; Huttenhower, C.; Geer, L. Y.; Coon, J. J.; Syka, J. E.; Bai, D. L.; Shabanowitz, J.; Burke, D. J.; Troyanskaya, O. G.; Hunt, D. F., Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc Natl Acad Sci U S A* **2007,** 104, (7), 2193-8.

36. Swaney, D. L.; Wenger, C. D.; Thomson, J. A.; Coon, J. J., Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A* **2009,** 106, (4), 995-1000.

37. Harvey, D. J., Proteomic analysis of glycosylation: structural determination of N- and O-linked glycans by mass spectrometry. *Expert Rev Proteomics* **2005,** 2, (1), 87-101.

38. Burlingame, A. L., Characterization of protein glycosylation by mass spectrometry. *Curr Opin Biotechnol* **1996,** 7, (1), 4-10.

39. Wuhrer, M.; Catalina, M. I.; Deelder, A. M.; Hokke, C. H., Glycoproteomics based on tandem mass spectrometry of glycopeptides. *J Chromatogr B Analyt Technol Biomed Life Sci* **2007,** 849, (1-2), 115-28.

40. Kelleher, N. L., Top-down proteomics. *Anal Chem* **2004,** 76, (11), 197A-203A.

41. Carr, S. A.; Huddleston, M. J.; Bean, M. F., Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Sci* **1993,** 2, (2), 183-96.

42. Huddleston, M. J.; Bean, M. F.; Carr, S. A., Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests. *Anal Chem* **1993,** 65, (7), 877-84.

43. Kieliszewski, M. J.; O'Neill, M.; Leykam, J.; Orlando, R., Tandem mass spectrometry and structural elucidation of glycopeptides from a hydroxyproline-rich plant cell wall glycoprotein indicate that contiguous hydroxyproline residues are the major sites of hydroxyproline O-arabinosylation. *J Biol Chem* **1995,** 270, (6), 2541-9.

44. Medzihradszky, K. F.; Gillece-Castro, B. L.; Settineri, C. A.; Townsend, R. R.; Masiarz, F. R.; Burlingame, A. L., Structure determination of O-linked glycopeptides by tandem mass spectrometry. *Biomed Environ Mass Spectrom* **1990,** 19, (12), 777-81.

45. Settineri, C. A.; Medzihradszky, K. F.; Masiarz, F. R.; Burlingame, A. L.; Chu, C.; George-Nascimento, C., Characterization of O-glycosylation sites in recombinant B-chain of platelet-derived growth factor expressed in yeast using liquid secondary ion mass spectrometry, tandem mass spectrometry and Edman sequence analysis. *Biomed Environ Mass Spectrom* **1990,** 19, (11), 665-76.

46. Odani, H.; Yamamoto, K.; Iwayama, S.; Iwase, H.; Takasaki, A.; Takahashi, K.; Fujita, Y.; Sugiyama, S.; Hiki, Y., Evaluation of the specific structures of IgA1 hinge glycopeptide in 30 IgA nephropathy patients by mass spectrometry. *J Nephrol* **2010,** 23, (1), 70-6.

47. Sullivan, B.; Addona, T. A.; Carr, S. A., Selective detection of glycopeptides on ion trap mass spectrometers. *Anal Chem* **2004,** 76, (11), 3112-8.

48. Temporini, C.; Perani, E.; Calleri, E.; Dolcini, L.; Lubda, D.; Caccialanza, G.; Massolini, G., Pronase-immobilized enzyme reactor: an approach for automation in glycoprotein analysis by LC/LC-ESI/MSn. *Anal Chem* **2007,** 79, (1), 355-63.

49. Wuhrer, M.; Koeleman, C. A.; Deelder, A. M., Hexose rearrangements upon fragmentation of N-glycopeptides and reductively aminated N-glycans. *Anal Chem* **2009,** 81, (11), 4422-32.

50. Zhang, L.; Reilly, J. P., Extracting both peptide sequence and glycan structural information by 157 nm photodissociation of N-linked glycopeptides. *J Proteome Res* **2009,** 8, (2), 734-42.

51. Catalina, M. I.; Koeleman, C. A.; Deelder, A. M.; Wuhrer, M., Electron transfer dissociation of N-glycopeptides: loss of the entire N-glycosylated asparagine side chain. *Rapid Commun Mass Spectrom* **2007,** 21, (6), 1053-61.

52. Demelbauer, U. M.; Zehl, M.; Plematl, A.; Allmaier, G.; Rizzi, A., Determination of glycopeptide structures by multistage mass spectrometry with low-energy collision-induced dissociation: comparison of electrospray ionization quadrupole ion trap and matrix-assisted laser desorption/ionization quadrupole ion trap reflectron time-of-flight approaches. *Rapid Commun Mass Spectrom* **2004,** 18, (14), 1575-82.

53. Sandra, K.; Devreese, B.; Van Beeumen, J.; Stals, I.; Claeyssens, M., The Q-Trap mass spectrometer, a novel tool in the study of protein glycosylation. *J Am Soc Mass Spectrom* **2004,** 15, (3), 413-23.

54. Hashii, N.; Kawasaki, N.; Itoh, S.; Nakajima, Y.; Harazono, A.; Kawanishi, T.; Yamaguchi, T., Identification of glycoproteins carrying a target glycan-motif by liquid chromatography/multiple-stage mass spectrometry: identification of Lewis x-conjugated glycoproteins in mouse kidney. *J Proteome Res* **2009,** 8, (7), 3415-29.

55. Peterman, S. M.; Mulholland, J. J., A novel approach for identification and characterization of glycoproteins using a hybrid linear ion trap/FT-ICR mass spectrometer. *J Am Soc Mass Spectrom* **2006,** 17, (2), 168-79.

56. Wang, Y.; Wu, S. L.; Hancock, W. S., Approaches to the study of N-linked glycoproteins in human plasma using lectin affinity chromatography and nano-HPLC coupled to electrospray linear ion trap--Fourier transform mass spectrometry. *Glycobiology* **2006,** 16, (6), 514-23.

57. Wang, Y.; Wu, S. L.; Hancock, W. S., Monitoring of glycoprotein products in cell culture lysates using lectin affinity chromatography and capillary HPLC coupled to electrospray linear ion trap-Fourier transform mass spectrometry (LTQ/FTMS). *Biotechnol Prog* **2006,** 22, (3), 873-80.

58. Deguchi, K.; Ito, H.; Baba, T.; Hirabayashi, A.; Nakagawa, H.; Fumoto, M.; Hinou, H.; Nishimura, S., Structural analysis of O-glycopeptides employing negative- and positive-ion multi-stage mass spectra obtained by collision-induced and electron-capture dissociations in linear ion trap time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **2007,** 21, (5), 691-8.

59. Ito, H.; Takegawa, Y.; Deguchi, K.; Nagai, S.; Nakagawa, H.; Shinohara, Y.; Nishimura, S., Direct structural assignment of neutral and sialylated N-glycans of

glycopeptides using collision-induced dissociation MSn spectral matching. *Rapid Commun Mass Spectrom* **2006,** 20, (23), 3557-65.

60. Harazono, A.; Kawasaki, N.; Itoh, S.; Hashii, N.; Ishii-Watabe, A.; Kawanishi, T.; Hayakawa, T., Site-specific N-glycosylation analysis of human plasma ceruloplasmin using liquid chromatography with electrospray ionization tandem mass spectrometry. *Anal Biochem* **2006,** 348, (2), 259-68.

61. Henriksson, H.; Denman, S. E.; Campuzano, I. D.; Ademark, P.; Master, E. R.; Teeri, T. T.; Brumer, H., 3rd, N-linked glycosylation of native and recombinant cauliflower xyloglucan endotransglycosylase 16A. *Biochem J* **2003,** 375, (Pt 1), 61-73.

62. Imre, T.; Schlosser, G.; Pocsfalvi, G.; Siciliano, R.; Molnar-Szollosi, E.; Kremmer, T.; Malorni, A.; Vekey, K., Glycosylation site analysis of human alpha-1-acid glycoprotein (AGP) by capillary liquid chromatography-electrospray mass spectrometry. *J Mass Spectrom* **2005,** 40, (11), 1472-83.

63. Nemeth, J. F.; Hochgesang, G. P., Jr.; Marnett, L. J.; Caprioli, R. M., Characterization of the glycosylation sites in cyclooxygenase-2 using mass spectrometry. *Biochemistry* **2001,** 40, (10), 3109-16.

64. Satomi, Y.; Shimonishi, Y.; Hase, T.; Takao, T., Site-specific carbohydrate profiling of human transferrin by nano-flow liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom* **2004,** 18, (24), 2983-8.

65. Satomi, Y.; Shimonishi, Y.; Takao, T., N-glycosylation at Asn(491) in the Asn-Xaa-Cys motif of human transferrin. *FEBS Lett* **2004,** 576, (1-2), 51-6.

66. Bykova, N. V.; Rampitsch, C.; Krokhin, O.; Standing, K. G.; Ens, W., Determination and characterization of site-specific N-glycosylation using MALDI-Qq-TOF tandem mass spectrometry: case study with a plant protease. *Anal Chem* **2006,** 78, (4), 1093-103.

67. Krokhin, O.; Ens, W.; Standing, K. G.; Wilkins, J.; Perreault, H., Site-specific N-glycosylation analysis: matrix-assisted laser desorption/ionization quadrupole-quadrupole time-of-flight tandem mass spectral signatures for recognition and identification of glycopeptides. *Rapid Commun Mass Spectrom* **2004,** 18, (18), 2020-30.

68. Krokhin, O. V.; Ens, W.; Standing, K. G., MALDI QqTOF MS combined with off-line HPLC for characterization of protein primary structure and post-translational modifications. *J Biomol Tech* **2005,** 16, (4), 429-40.

69. Kubota, K.; Sato, Y.; Suzuki, Y.; Goto-Inoue, N.; Toda, T.; Suzuki, M.; Hisanaga, S.; Suzuki, A.; Endo, T., Analysis of glycopeptides using lectin affinity chromatography with MALDI-TOF mass spectrometry. *Anal Chem* **2008,** 80, (10), 3693-8.

70. Suzuki, Y.; Suzuki, M.; Nakahara, Y.; Ito, Y.; Ito, E.; Goto, N.; Miseki, K.; Iida, J.; Suzuki, A., Structural characterization of glycopeptides by N-terminal protein ladder sequencing. *Anal Chem* **2006,** 78, (7), 2239-43.

71. Irungu, J.; Go, E. P.; Zhang, Y.; Dalpathado, D. S.; Liao, H. X.; Haynes, B. F.; Desaire, H., Comparison of HPLC/ESI-FTICR MS versus MALDI-TOF/TOF MS for glycopeptide analysis of a highly glycosylated HIV envelope glycoprotein. *J Am Soc Mass Spectrom* **2008,** 19, (8), 1209-20.

72. Sparbier, K.; Asperger, A.; Resemann, A.; Kessler, I.; Koch, S.; Wenzel, T.; Stein, G.; Vorwerg, L.; Suckau, D.; Kostrzewa, M., Analysis of glycoproteins in human serum by means of glycospecific magnetic bead separation and LC-MALDI-TOF/TOF analysis with automated glycopeptide detection. *J Biomol Tech* **2007,** 18, (4), 252-8.

73. Wuhrer, M.; Hokke, C. H.; Deelder, A. M., Glycopeptide analysis by matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry reveals novel features of horseradish peroxidase glycosylation. *Rapid Commun Mass Spectrom* **2004,** 18, (15), 1741-8.

74. Domon, B.; Costello, C. E., A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal* **1988,** 5, (4), 397-409.

75. Takemori, N.; Komori, N.; Matsumoto, H., Highly sensitive multistage mass spectrometry enables small-scale analysis of protein glycosylation from two-dimensional polyacrylamide gels. *Electrophoresis* **2006,** 27, (7), 1394-406.

76. Nishimura, S.; Niikura, K.; Kurogochi, M.; Matsushita, T.; Fumoto, M.; Hinou, H.; Kamitani, R.; Nakagawa, H.; Deguchi, K.; Miura, N.; Monde, K.; Kondo, H., High-throughput protein glycomics: combined use of chemoselective glycoblotting and MALDI-TOF/TOF mass spectrometry. *Angew Chem Int Ed Engl* **2004,** 44, (1), 91-6.

77. Kurogochi, M.; Nishimura, S., Structural characterization of N-glycopeptides by matrix-dependent selective fragmentation of MALDI-TOF/TOF tandem mass spectrometry. *Anal Chem* **2004,** 76, (20), 6097-101.

78. Kurogochi, M.; Matsushita, T.; Nishimura, S., Post-translational modifications on proteins: facile and efficient procedure for the identification of O-glycosylation sites by MALDI-LIFT-TOF/TOF mass spectrometry. *Angew Chem Int Ed Engl* **2004,** 43, (31), 4071-5.

79. Wuhrer, M.; Balog, C. I.; Koeleman, C. A.; Deelder, A. M.; Hokke, C. H., New features of site-specific horseradish peroxidase (HRP) glycosylation uncovered by nano-LC-MS with repeated ion-isolation/fragmentation cycles. *Biochim Biophys Acta* **2005,** 1723, (1-3), 229-39.

80. Adamson, J. T.; Hakansson, K., Infrared multiphoton dissociation and electron capture dissociation of high-mannose type glycopeptides. *J Proteome Res* **2006,** 5, (3), 493-501.

81. Hakansson, K.; Chalmers, M. J.; Quinn, J. P.; McFarland, M. A.; Hendrickson, C. L.; Marshall, A. G., Combined electron capture and infrared multiphoton dissociation for multistage MS/MS in a Fourier transform ion cyclotron resonance mass spectrometer. *Anal Chem* **2003,** 75, (13), 3256-62.

82. Hakansson, K.; Cooper, H. J.; Emmett, M. R.; Costello, C. E.; Marshall, A. G.; Nilsson, C. L., Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptic to yield complementary sequence information. *Anal Chem* **2001,** 73, (18), 4530-6.

83. Seipert, R. R.; Dodds, E. D.; Clowers, B. H.; Beecroft, S. M.; German, J. B.; Lebrilla, C. B., Factors that influence fragmentation behavior of N-linked glycopeptide ions. *Anal Chem* **2008,** 80, (10), 3684-92.

84. Haselmann, K. F.; Budnik, B. A.; Olsen, J. V.; Nielsen, M. L.; Reis, C. A.; Clausen, H.; Johnsen, A. H.; Zubarev, R. A., Advantages of external accumulation for electron capture dissociation in Fourier transform mass spectrometry. *Anal Chem* **2001,** 73, (13), 2998-3005.

85. Kjeldsen, F.; Haselmann, K. F.; Budnik, B. A.; Sorensen, E. S.; Zubarev, R. A., Complete characterization of posttranslational modification sites in the bovine milk protein PP3 by tandem mass spectrometry with electron capture dissociation as the last stage. *Anal Chem* **2003,** 75, (10), 2355-61.

86. Mirgorodskaya, E.; Roepstorff, P.; Zubarev, R. A., Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal Chem* **1999,** 71, (20), 4431-6.

87. Mormann, M.; Paulsen, H.; Peter-Katalinic, J., Electron capture dissociation of O-glycosylated peptides: radical site-induced fragmentation of glycosidic bonds. *Eur J Mass Spectrom (Chichester, Eng)* **2005,** 11, (5), 497-511.

88. Renfrow, M. B.; Cooper, H. J.; Tomana, M.; Kulhavy, R.; Hiki, Y.; Toma, K.; Emmett, M. R.; Mestecky, J.; Marshall, A. G.; Novak, J., Determination of aberrant O-glycosylation in the IgA1 hinge region by electron capture dissociation fourier transform-ion cyclotron resonance mass spectrometry. *J Biol Chem* **2005,** 280, (19), 19136-45.

89. Alley, W. R., Jr.; Mechref, Y.; Novotny, M. V., Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. *Rapid Commun Mass Spectrom* **2009,** 23, (1), 161-70.

90. Darula, Z.; Medzihradszky, K. F., Affinity enrichment and characterization of mucin core-1 type glycopeptides from bovine serum. *Mol Cell Proteomics* **2009,** 8, (11), 2515-26.

91. Snovida, S. I.; Bodnar, E. D.; Viner, R.; Saba, J.; Perreault, H., A simple cellulose column procedure for selective enrichment of glycopeptides and characterization by nano LC coupled with electron-transfer and high-energy collisional-dissociation tandem mass spectrometry. *Carbohydr Res* **2010,** 345, (6), 792-801.

92. Dalpathado, D. S.; Irungu, J.; Go, E. P.; Butnev, V. Y.; Norton, K.; Bousfield, G. R.; Desaire, H., Comparative glycomics of the glycoprotein follicle stimulating hormone: glycopeptide analysis of isolates from two mammalian species. *Biochemistry* **2006,** 45, (28), 8665-73.

93. Geng, M.; Zhang, X.; Bina, M.; Regnier, F., Proteomics of glycoproteins based on affinity selection of glycopeptides from tryptic digests. *J Chromatogr B Biomed Sci Appl* **2001,** 752, (2), 293-306.

94. Irungu, J.; Dalpathado, D. S.; Go, E. P.; Jiang, H.; Ha, H. V.; Bousfield, G. R.; Desaire, H., Method for characterizing sulfated glycoproteins in a glycosylation site-specific fashion, using ion pairing and tandem mass spectrometry. *Anal Chem* **2006,** 78, (4), 1181-90.

95. Jiang, H.; Desaire, H.; Butnev, V. Y.; Bousfield, G. R., Glycoprotein profiling by electrospray mass spectrometry. *J Am Soc Mass Spectrom* **2004,** 15, (5), 750-8.

96. Liu, T.; Li, J. D.; Zeng, R.; Shao, X. X.; Wang, K. Y.; Xia, Q. C., Capillary electrophoresis-electrospray mass spectrometry for the characterization of high-mannose-type N-glycosylation and differential oxidation in glycoproteins by charge reversal and protease/glycosidase digestion. *Anal Chem* **2001,** 73, (24), 5875-85.

97. Wang, F.; Nakouzi, A.; Angeletti, R. H.; Casadevall, A., Site-specific characterization of the N-linked oligosaccharides of a murine immunoglobulin M by high-performance liquid chromatography/electrospray mass spectrometry. *Anal Biochem* **2003,** 314, (2), 266-80.

98. Sihlbom, C.; van Dijk Hard, I.; Lidell, M. E.; Noll, T.; Hansson, G. C.; Backstrom, M., Localization of O-glycans in MUC1 glycoproteins using electron-capture dissociation fragmentation mass spectrometry. *Glycobiology* **2009,** 19, (4), 375-81.

99. Renfrow, M. B.; Mackay, C. L.; Chalmers, M. J.; Julian, B. A.; Mestecky, J.; Kilian, M.; Poulsen, K.; Emmett, M. R.; Marshall, A. G.; Novak, J., Analysis of O-glycan heterogeneity in IgA1 myeloma proteins by Fourier transform ion cyclotron resonance mass spectrometry: implications for IgA nephropathy. *Anal Bioanal Chem* **2007,** 389, (5), 1397-407.

100. Clowers, B. H.; Dodds, E. D.; Seipert, R. R.; Lebrilla, C. B., Site determination of protein glycosylation based on digestion with immobilized nonspecific proteases and Fourier transform ion cyclotron resonance mass spectrometry. *J Proteome Res* **2007,** 6, (10), 4032-40.

101. Vosseller, K.; Trinidad, J. C.; Chalkley, R. J.; Specht, C. G.; Thalhammer, A.; Lynn, A. J.; Snedecor, J. O.; Guan, S.; Medzihradszky, K. F.; Maltby, D. A.; Schoepfer, R.; Burlingame, A. L., O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. *Mol Cell Proteomics* **2006,** 5, (5), 923-34.

102. Muller, S.; Alving, K.; Peter-Katalinic, J.; Zachara, N.; Gooley, A. A.; Hanisch, F. G., High density O-glycosylation on tandem repeat peptide from secretory MUC1 of T47D breast cancer cells. *J Biol Chem* **1999,** 274, (26), 18165-72.

103. Muller, S.; Goletz, S.; Packer, N.; Gooley, A.; Lawson, A. M.; Hanisch, F. G., Localization of O-glycosylation sites on glycopeptide fragments from lactation-associated MUC1. All putative sites within the tandem repeat are glycosylation targets in vivo. *J Biol Chem* **1997,** 272, (40), 24780-93.

104. Macek, B.; Hofsteenge, J.; Peter-Katalinic, J., Direct determination of glycosylation sites in O-fucosylated glycopeptides using nano-electrospray quadrupole time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **2001,** 15, (10), 771-7.

105. Hanisch, F. G.; Green, B. N.; Bateman, R.; Peter-Katalinic, J., Localization of O-glycosylation sites of MUC1 tandem repeats by QTOF ESI mass spectrometry. *J Mass Spectrom* **1998,** 33, (4), 358-62.

106. Chalkley, R. J.; Burlingame, A. L., Identification of GlcNAcylation sites of peptides and alpha-crystallin using Q-TOF mass spectrometry. *J Am Soc Mass Spectrom* **2001,** 12, (10), 1106-13.

107. Chalabi, S.; Panico, M.; Sutton-Smith, M.; Haslam, S. M.; Patankar, M. S.; Lattanzio, F. A.; Morris, H. R.; Clark, G. F.; Dell, A., Differential O-glycosylation of a conserved domain expressed in murine and human ZP3. *Biochemistry* **2006,** 45, (2), 637-47.

108. Joenvaara, S.; Ritamo, I.; Peltoniemi, H.; Renkonen, R., N-glycoproteomics - an automated workflow approach. *Glycobiology* **2008,** 18, (4), 339-49.

109. Atwood, J. A., 3rd; Minning, T.; Ludolf, F.; Nuccio, A.; Weatherly, D. B.; Alvarez-Manilla, G.; Tarleton, R.; Orlando, R., Glycoproteomics of Trypanosoma cruzi trypomastigotes using subcellular fractionation, lectin affinity, and stable isotope labeling. *J Proteome Res* **2006,** 5, (12), 3376-84.

110. Kaji, H.; Saito, H.; Yamauchi, Y.; Shinkawa, T.; Taoka, M.; Hirabayashi, J.; Kasai, K.; Takahashi, N.; Isobe, T., Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol* **2003,** 21, (6), 667-72.

111. Xiong, L.; Regnier, F. E., Use of a lectin affinity selector in the search for unusual glycosylation in proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **2002,** 782, (1-2), 405-18.

112. Edge, A. S., Deglycosylation of glycoproteins with trifluoromethanesulphonic acid: elucidation of molecular structure and function. *Biochem J* **2003,** 376, (Pt 2), 339-50.

113. Gerken, T. A.; Gupta, R.; Jentoft, N., A novel approach for chemically deglycosylating O-linked glycoproteins. The deglycosylation of submaxillary and respiratory mucins. *Biochemistry* **1992,** 31, (3), 639-48.

114. Greis, K. D.; Hayes, B. K.; Comer, F. I.; Kirk, M.; Barnes, S.; Lowary, T. L.; Hart, G. W., Selective detection and site-analysis of O-GlcNAc-modified glycopeptides by beta-elimination and tandem electrospray mass spectrometry. *Anal Biochem* **1996,** 234, (1), 38-49.

115. Durham, M.; Regnier, F. E., Targeted glycoproteomics: serial lectin affinity chromatography in the selection of O-glycosylation sites on proteins from the human blood proteome. *J Chromatogr A* **2006,** 1132, (1-2), 165-73.

116. Hong, J. C.; Kim, Y. S., Alkali-catalyzed beta-elimination of periodate-oxidized glycans: a novel method of chemical deglycosylation of mucin gene products in paraffin embedded sections. *Glycoconj J* **2000,** 17, (10), 691-703.

117. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **2002,** 1, (5), 376-86.

118. Ong, S. E.; Mann, M., A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc* **2006,** 1, (6), 2650-60.

119. Miyagi, M.; Rao, K. C., Proteolytic 18O-labeling strategies for quantitative proteomics. *Mass Spectrom Rev* **2007,** 26, (1), 121-36.

120. Staes, A.; Demol, H.; Van Damme, J.; Martens, L.; Vandekerckhove, J.; Gevaert, K., Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18. *J Proteome Res* **2004,** 3, (4), 786-91.

121. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **1999,** 17, (10), 994-9.

122. Schmidt, A.; Kellermann, J.; Lottspeich, F., A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* **2005,** 5, (1), 4-15.

123. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **2003,** 100, (12), 6940-5.

124. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004,** 3, (12), 1154-69.

125. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **2003,** 75, (8), 1895-904.

126. Zhang, G.; Neubert, T. A., Automated comparative proteomics based on multiplex tandem mass spectrometry and stable isotope labeling. *Mol Cell Proteomics* **2006,** 5, (2), 401-11.

127. Venable, J. D.; Dong, M. Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R., Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **2004,** 1, (1), 39-45.

128. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004,** 76, (14), 4193-201.

129. Wang, N.; Li, L., Exploring the precursor ion exclusion feature of liquid chromatography-electrospray ionization quadrupole time-of-flight mass spectrometry for improving protein identification in shotgun proteome analysis. *Anal Chem* **2008,** 80, (12), 4696-710.

130. Asara, J. M.; Christofk, H. R.; Freimark, L. M.; Cantley, L. C., A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *Proteomics* **2008,** 8, (5), 994-9.

131. Rappsilber, J.; Ryder, U.; Lamond, A. I.; Mann, M., Large-scale proteomic analysis of the human spliceosome. *Genome Res* **2002,** 12, (8), 1231-45.

132. Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M., Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **2005,** 4, (9), 1265-72.

133. Lu, P.; Vogel, C.; Wang, R.; Yao, X.; Marcotte, E. M., Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **2007,** 25, (1), 117-24.

134. Piques, M.; Schulze, W. X.; Hohne, M.; Usadel, B.; Gibon, Y.; Rohwer, J.; Stitt, M., Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. *Mol Syst Biol* **2009,** 5, 314.

135. Zhang, X.; Julien-David, D.; Miesch, M.; Raul, F.; Geoffroy, P.; Aoude-Werner, D.; Ennahar, S.; Marchioni, E., Quantitative analysis of beta-sitosterol oxides induced in vegetable oils by natural sunlight, artificially generated light, and irradiation. *J Agric Food Chem* **2006,** 54, (15), 5410-5.

136. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* **2003,** 75, (18), 4818-26.

137. Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S., Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem* **2005,** 77, (7), 2187-200.

138. Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J., Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **2006,** 5, (1), 144-56.

139. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **2008,** 4, 222.

CHAPTER 2

THE HUMAN EMBRYONIC STEM CELL PROTEOME REVEALED BY MULTIDIMENSIONAL FRACTIONATION FOLLOWED BY TANDEM MASS SPECTROMETRY[1]

[1] Peng Zhao, Thomas C. Schulz, D. Brent Weatherly, Allan J. Robins, Lance Wells
To be submitted to *Journal of Proteome Research*.

ABSTRACT

Human embryonic stem cells (hESCs) have received considerable attention due to their therapeutic potential and their usefulness in understanding early development and cell fate commitment. In order to appreciate the unique properties of these pluripotent, self-renewing cells, the proteins made, the post-translational modifications that occur, and the proteins targeted for secretion need to be elucidated. Further, these cells need to be characterized when grown under the conditions that would be used for any therapeutic purposes, namely defined media conditions without feeder layers or serum. Thus, we have performed an in-depth multidimensional fractionation followed by LC-MS/MS analysis of the hES cell line BG02 harvested from defined media conditions. Following triplicate analysis of each generated peptide fraction, we were able to assign more than 3,000 proteins with less than 1% false discovery rate. Clustering of these identified proteins generated a plethora of information regarding biological pathways that are operating in these stem cells. The analysis of these more than 300 hundred tandem mass spectrometry experiments also allowed us to identify nearly 500 phosphorylation sites and 68 sites of O-GlcNAc modification with the same high confidence. Analysis of the phosphorylation sites allowed us to infer what kinases are likely active in these cells, and inhibition assays were performed on a subset of these kinases to establish a functional role for the enzymes in these cells. Since these cells were grown under defined media conditions, we were able to characterize the secreted proteome of the BG02 cell line and identify more than 100 secreted proteins that likely play a role in extracellular matrix formation and remodeling as well as autocrine signaling for self-renewal and maintenance of the undifferentiated state. Finally, by performing in-depth analysis in

triplicate, spectral counts with standard deviations were obtained for many of these proteins and post-translationally modified peptides that will allow us to perform relative quantitative analysis between these cells and any derived cell type.

INTRODUCTION

Embryonic stem cells (ESCs), which are derived from blastocysts, are pluripotent, having the theoretical potential to become most specialized cell types. Studies using ESCs not only provide basic information about early development and cell fate decisions but may also serve as a potential source for the creation of large populations of a defined cell type. The latter subject raises possibilities of cell-based therapies to treat disease, often referred to as regenerative medicine. Data now shows that human embryonic stem cells can initiate lineage-specific differentiation programs of many tissue and cell types *in vitro*[1-2]. For example, the creation of primitive endoderm and mesoderm under defined conditions raises the realistic possibility for producing cardiac progenitors for heart tissue regeneration in cardiovascular disease therapies[3].

Given the likely importance of these cells in basic and translational science, several genome and transcriptome studies have been performed. Likewise, several proteomic efforts have elucidated some of the proteins expressed by these cells. The most extensive studies on the proteome of undifferentiated hESCs identified 1775 proteins[4] using FT-ICR-MS/MS, however, they didn't discover any post-translationally modified proteins. Other studies on pluripotent hESCs with less depth of proteomic coverage, 685[5] and 589[6] proteins assigned respectively, did identify 251[5] and 76[6] post-translationally modified proteins using MALDI-TOF-MS/MS[5] and immunoblotting[6].

Multidimensional chromatography coupled with mass spectrometry-based protein assignment technology has dramatically increased resolution and loading capacity for identifying large numbers of proteins with a wide dynamic range of abundances, especially for proteins with post-translational modifications such as phosphorylation and

glycosylation[7-10]. Here we describe a large-scale proteomic study of human embryonic stem cells (BG02) grown under defined conditions without serum or feeder layers using multidimensional chromatography coupled with tandem mass spectrometry. We applied offline SCX and RP HPLC to provide more in-depth peptide separation, and LC-MS/MS techniques for protein assignment (Figure 2-1). We also investigated the post-translational modification of the hES cell proteome, particularly sites of phosphorylation and O-GlcNAc modification. Furthermore, since the cells were cultured in defined media, we were able to explore the secreted proteome of these cells. Functional classification analysis of this large, high-confidence hES cell proteome with post-translational modifications elucidates many of the biological features of this pluripotent cell type.

EXPERIMENTAL PROCEDURE

Culture of human embryonic stem cells

Karyotypically normal cultures of the NIH-registered BG02 hES cell line were used throughout the experiments. BG02 cells were grown in DC-HAIF defined medium[11], which consisted of DMEM/F12 (Invitrogen), 2% fatty acid-free Cohn's fraction V BSA (Serologicals), 1x Nonessential amino acids, 50 U/ml Penicillin, 50 µg/ml Streptomycin, 50 µg/ml Ascorbic Acid, 10 µg/ml Transferrin, 0.1 mM β-Mercaptoethanol (all from Invitrogen), 1x Trace Elements A, B & C (Mediatech), 10 ng/ml HRG1β (Peprotech), 10 ng/ml ActA (R&D Systems), 200 ng/ml LR^3-IGF1 (JRH Biosciences), and 8 ng/ml FGF2 (Sigma or R&D Systems). Cultures were expanded by passaging with Accutase (Innovative Cell Technologies) and plating on tissue culture flasks coated with growth factor-reduced matrigel (BD Biosciences) diluted 1:200, as

described[12]. To create samples for proteomics, near-confluent cultures were washed extensively with PBS and grown for 24 hours in DC-HAIF without BSA. Cultures were harvested with Accutase, washed with PBS, pelleted and snap frozen. Initial experiments were performed on the undifferentiated population, with an estimated 5 x 10^7 provided for each sample run.

Subcellular fractionation, protein precipitation and digestion

Cells were fractionated into 4 subcellular proteomes using ProteoExtract® Subcellular Proteome Extraction Kit (Calbiochem): membrane/organelle, nucleus, cytosol, and cytoskeleton. Specifically, the pellet was resuspended and rotated at 4 °C for 10 min in ice-cold Extraction Buffer I (Calbiochem) with protease and phosphotase inhibitors (Sigma). The suspension was then centrifuged for 10 min at 800 G at 4 °C and the supernatant was collected as cytosol fraction. The remaining pellet was resuspended and rotated at 4 °C for 30 min in ice-cold Extraction Buffer II (Calbiochem) with protease and phosphotase inhibitors (Sigma). The suspension was then centrifuged for 10 min at 5500 G at 4 °C and the supernatant was collected as membrane/organelle fraction. The remaining pellet was resuspended and rotated at 4 °C for 10 min in ice-cold Extraction Buffer III and benzonase nuclease (Calbiochem) with protease and phosphotase inhibitors (Sigma). The suspension was then centrifuged for 10 min at 7000 G at 4 °C and the supernatant was collected as nucleus fraction. The remaining pellet was resuspended in ice-cold Extraction Buffer IV (Calbiochem) with protease and phosphotase inhibitors (Sigma) and then collected as cytoskeleton fraction.

Proteins from 4 sub-proteomes were then precipitated using ProteoExtract® Protein Precipitation Kit (Calbiochem). Each fraction was incubated in Precipitation

Agent (Calbiochem) at -20 °C for 1 hr and then centrifuged for 10 min at 10000 G at 4 °C. The resulting pellet was washed 3 times with ice-cold Wash Solution (Calbiochem) and centrifuged for 2 min at 10000 G at 4 °C. The pellet from each fraction was collected, dried at room temperature for 15 min and resuspended in the mixture of 8 M urea, 10 mM DTT, 40 mM NH_4CO_3 and incubated at 37 °C for 1 hr. After cooling to room temperature, the suspension was incubated for 45 min in dark with 55 mM iodoacetamide dissolved in 40 mM NH_4CO_3. After that, each fraction was digested overnight at 37 °C using sequencing grade trypsin (Promega).

Offline strong cation exchange liquid chromatography (1st dimension SCX-LC)

Peptides generated from digestion were separated by offline strong cation exchange liquid chromatography (SCX-LC) using an Agilent 1100 series HPLC system (Agilent Technologies). Solvent A (5 mM KH_2PO_4/30% acetonitrile, pH 2.7), solvent B (solvent A with 350 mM KCl), and solvent C (0.1 M Tris/0.5 M KCl, pH 7.0) were used to develop a salt gradient consisting of 5 min at 100% solvent A, 48 min gradient at variable slope to 100% solvent B, 12 min at 100% solvent B, 15 min to 100% solvent C, and 10 min to 100% solvent A. Dried peptides of each sub-proteome were resuspended in 50 µl of SCX solvent A and separated on a 2.1 x 100 mm Polysulfoethyl A column (PolyLC) at a flow rate of 150 µl/min. Fractions were collected every 2 min, and then combined into 5 fractions, desalted and dried.

Offline reverse phase liquid chromatography (2nd dimension RP-LC)

Each fraction generated by SCX-LC (5×4 fractions in total) was separated by reverse phase liquid chromatography (RP-LC) using the same Agilent 1100 series HPLC system (Agilent Technologies). Solvent A (0.1% Trifluoroacetic acid) and solvent B

(0.085% Trifluoroacetic acid/80% acetonitrile) were used to develop a linear gradient consisting of 5 min at 95% solvent A, 60 min gradient at variable slope to 95% solvent B, 3 min at 95% solvent B, 1.5 min to 95% solvent A, and 4.5 min at 95% solvent A. Dried peptides of each sub-proteome were dissolved in 20 µl of RP solvent A and separated on a 2.1 x 250 mm silica-based C18 column (VYDAC) at a flow rate of 100 µl/min. Fractions were collected every 4 min, and then combined into 5 fractions (F1, 15-32%; F2, 32-40%; F3, 40-45%; F4, 45-55%; and F5, 55-85%), desalted and dried.

Reverse phase nanoLC-MS/MS analysis (3rd dimension RP-nLC-MS/MS)

Dried peptides from each fraction generated by RP-LC (25×4 in total) were resuspended in 0.5 µl of solvent B (0.1% Formic acid/80% acetonenitrile) and 19.5 µl of solvent A (0.1% Formic acid) and loaded on a 75 µm x 85 mm C18 reverse phase column (packed in house, YMC GEL ODS-AQ120ÅS-5, Waters) by nitrogen bomb. Peptides were eluted directly into the nanospray source of a linear ion trap (Thermo Finnigan LTQ™) with a 140-min linear gradient consisting of 5-100% solvent B over 90-95 min at a flow rate of ~250 nl/min. In order to optimize the separation of peptides eluted into the mass spectrometer, gradients were expanded over a 70-min period in the appropriate region corresponding to each fraction collected from the previous offline RP-LC separation (F1, 4-30%; F2, 9-35%; F3, 15-42%; F4, 20-55%; and F5, 28-85%). The spray voltage was set to 2.0 kV and the temperature of the heated capillary was set to 210 °C. Full scan MS spectra were acquired from m/z 300 to 2000 followed by 8 MS/MS events of the most intense ions. Subsequent MS/MS/MS events were triggered by the detection of neutral losses of phosphoric acid (98, 49, or 32.67 amu) or GlcNAc (203.08, 101.55, or 67.67 amu) between precursor ions and the resulting 3 most intense product ions (Figure

2-2). A dynamic exclusion window was applied which prevents the same m/z value from being selected for 12 seconds after its acquisition. All 100 sub-fractions were analyzed in 3 technical replicates. Data were automatically acquired using Xcalibur® (ver. 2.0.7, Thermo Fisher Scientific). It has been shown that for protein identification in mixture by tandem mass spectrometry, certain number of repeated experiments is required to reach a reasonable completeness (all predicted proteins in the mixture being identified). From statistical results, 3 replicates in MS/MS experiment can discover approximately 95% of all predicted proteins in samples with relatively high complexity.[13]

Single dimensional LC-MS/MS experiment

In order to compare the depths of protein identification between single dimensional and multidimensional experiments, approximately 1% of the peptides generated from previous tryptic digestion described above were directly analyzed by online nanospray LC coupled with linear ion trap (Thermo Finnigan LTQ™) without prior offline chromatography separation (SCX or RP-LC). The online RP-LC gradient used for peptides of each subcellular fractions was a 160-min linear gradient consisting of 5-100% solvent B over 100 min at a flow rate of ~250 nl/min. The parameters of the instrument set-up and analysis method were as described above.

Secreted proteome of hESCs

The secreted proteins were digested and the resulting peptides were separated by offline SCX chromatography as described above. Dried peptides from each fraction (5 fractions in total) were resuspended in 0.5 µl of solvent B (0.1% Formic acid/80% acetonenitrile) and 19.5 µl of solvent A (0.1% Formic acid) and loaded on a 75 µm x 105 mm C18 reverse phase column (packed in house, YMC GEL ODS-AQ120ÅS-5, Waters)

by nitrogen bomb. Peptides were eluted directly into the nanospray source of an LTQ Orbitrap XL™ (Thermo Scientific) with a 160-min linear gradient consisting of 5-100% solvent B over 100 min at a flow rate of ~250 nl/min. The spray voltage was set to 2.0 kV and the temperature of the heated capillary was set to 200 °C. Full scan MS spectra were acquired from m/z 300 to 2000 with a resolution of 60000 at m/z 400 after accumulation of 1000000 ions (mass accuracy < 2 ppm). MS/MS events were triggered by the 6 most intense ions from the preview of full scan and a dynamic exclusion window was applied which prevents the same m/z value from being selected for 6 seconds after its acquisition. All 5 sub-fractions were analyzed in 3 technical replicates. Data were acquired using Xcalibur® (ver. 2.0.7, Thermo Fisher Scientific).

Data analysis

For protein identification, MS/MS and MS/MS/MS data were searched against Swiss-Prot human proteome database (32876 entries, Aug. 13, 2007) using SEQUEST (Bioworks 3.3, Thermo Fisher Scientific) with the following settings: 1000-ppm (10-ppm for secreted proteome data) and 0.5-Da deviation were set for precursor and fragment masses, respectively; trypsin was specified as the enzyme; only fully tryptic peptide identifications were retained; a maximum of 3 missed cleavage sites, 3 differential amino acids per modification and 3 differential modifications per peptide were allowed; oxidized methionine (+15.99 amu), carbamidomethylcysteine (+57.02 amu), phosphorylated serine/threonine/tyrosine (+79.97 amu), and O-GlcNAc modified serine/threonine (203.08 amu) were set as differential modifications. All of the raw spectra were searched against both normal and reversed database under the same parameters, and all of the output files from SEQUEST search were filtered and grouped

by different subcellular fractions and replicates in ProteoIQ™. The cutoff value of peptides was set to an Xcorr of 0.5 and the minimum peptide length was set to 4 amino acids. False discovery rate was set to 1% at protein level and peptides matched to multiple proteins were excluded for protein quantification. ProteoIQ™ is a software developed by Weatherly et al.[14] based on the algorithm PROVALT which extracts peptides matches from multiple search results files, eliminates peptide redundancy, and clusters peptides to their corresponding proteins. PROVALT compares the score (Xcorr, deltaCn) distributions obtained from searching the normal and random databases to calculate protein false discovery rates associated with each score threshold so that it can identify as many real proteins as possible while encountering a minimal number of false-positive protein identifications. The filtered and combined result was submitted to Ingenuity Pathway Analysis (IPA, www.ingenuity.com)[15-16] for function and pathway classification. IPA is built upon a huge foundation of scientific evidence, manually curated from hundreds of thousands of journal articles, textbooks, and other data sources, and it allows the researcher to explore molecular, chemical, gene, protein and mRNA interactions, create custom molecular pathways, view and modify metabolic, signaling, and toxicological canonical pathways, each with underlying experimental literature evidence.

Real-time electronic sensing of hESCs

Small molecule compounds were purchased from Sigma, in the LOPAC array (Sigma, LO1280). Individual compounds and Sigma catalog numbers are: CK2 Inhibitor 2 (C7367), TBBz (T6951), rac-2-Ethoxy-3-hexadecanamido-1-propylphosphocholine (E3645), rac-2-Ethoxy-3-octadecanamido-1-propylphosphocholine (E3770), 1-(5-

Isoquinolinylsulfonyl)-3-methylpiperazine dihydrochloride (I6391), 1-(5-Isoquinolinylsulfonyl)-2-methylpiperazine dihydrochloride (I7016), Chelerythrine chloride (C2932), Rottlerin (R5648), Oleic Acid (O1008), Phorbol 12-myristate 13-acetate (P8139), Palmitoyl-DL-Carnitine chloride (P4509). The effects of small molecule inhibitors on hESCs were assayed using the RT-CES system (ACEA Biosciences). RT-CES is a label-free, real time impedance-monitoring device for unbiased tracking of dynamic changes in live cell populations[17]. Changes to cell conductivity, shape, migration, adhesion, proliferation, apoptosis, and other characteristics, can be followed. BG02 cells were harvested with Accutase and 10^4 cells were plated in DC-HAIF medium into matrigel-coated wells of 16-well glass slides embedded with microsensors. The cultures were placed in the RT-CES reader station under standard humidified incubator conditions at 37 ˚C with 5% CO_2. The media was changed every 24 hours and impedance was measured every 15 minutes for 5 days. Compounds were added at 10 μM, 48 hours after plating. Triplicate wells were averaged for each compound. The DMSO carrier control was averaged from six wells. The impedance traces for all wells were normalized at the time point immediately prior to addition of compounds. The relative change in electrical impedance was plotted as Normalized Cell Index versus Time.

RESULTS

Assignment of hES cell proteome

Proteins were extracted from BG02 hESCs and separated into 4 subcellular proteomes. Tryptic digests of each sub-proteome were separated by a 90-min offline SCX-LC experiment and combined into 5 fractions (20 fractions in total). 5 SCX fractions of each sub proteome were then separated by 90-min offline RP-LC and

combined into 5 fractions (100 fractions in total). All 100 fractions were analyzed in 140-min capillary RP-LC coupled online to a linear ion trap mass spectrometer in triplicate for peptide identification (300 LC-MS/MS experiments in total) (Figure 2-1). Output files from SEQUEST search in Swiss Prot human database were filtered and assigned into protein groups by ProteoIQ™. The false discovery rate of the dataset was <1% at the protein level as determined from parallel analysis using Swiss Prot database with all protein sequences reversed. After eliminating duplicates and contaminants, the non-redundant combined dataset consists of 3189 proteins identified by ~24,000 peptides of which more than 77% (2469/3189) proteins were identified by more than one peptide (Table 2-1). Thus, our MS/MS analysis combined with stringent cutoff filters in ProteoIQ™ resulted in an in-depth, high-confidence assigned proteome. Additionally, since the LC-MS/MS analysis was performed in technical triplicates, 2604, 2245, and 2391 proteins, respectively, were identified in each replicate with approximately 65% overlap between any two of the triplicate (Figure 2-3A). Furthermore, 860 (26.97%) proteins were observed in only one of the triplicate, 607 (19.03%) were observed in two, and the majority of the proteins (1722, 54.00%) were identified at <1% FDR in all three (Figure 2-3B). The lack of 100% or close-to-100% reproducibility can be explained by the incomplete sampling of data-dependent acquisition by tandem mass spectrometer.[13]

Subcellular localization of identified proteins

Subcellular localization of all the identified proteins were predicted by PSORT (wolfpsort.org), which is an algorithm based on the full-length amino acid sequences of proteins, and subsequently confirmed by Ingenuity knowledge base. Proteins that didn't show consistent results from PSORT and Ingenuity were manually searched against

Human Protein Reference Database (www.hprd.org) or Gene Ontology web source (www.geneontology.org) Proteins that didn't show any consistency in four searches were classified as unknown proteins.

From Figure 2-4, most of the proteins in hESCs are localized in nucleus (40%) and cytoplasm (35%). Interestingly, 2% of them (74) were identified as extracellular proteins, likely reflecting proteins that were being processed through the secretory pathway, recycled, or associated with the plasma membrane.

Biological functions of identified proteins

Over 65% (2132/3189) of the identified proteins were able to be classified into 62 biological functions involving 804 different biological processes by IPA. The top 5 assigned biological functions are: RNA post-transcriptional modification, protein synthesis, molecular transport, protein trafficking, and post-translational modification. Approximately 75% (2367/3189) of the identified proteins were assigned into 100 biological networks by IPA and over 65% (2132/3189) of the identified proteins were involved in 93 metabolic and signaling pathways by IPA. The top 5 involved canonical pathways are: aminoacyl-tRNA biosynthesis, protein ubiquitination pathway, lysine degradation, valine, leucine and isoleucine degradation, and tight junction signaling.

Assignment of the phosphorylated proteome within hESCs

Based on the dataset of identified proteins, all the peptides that had shown additional mass of one or more phosphoric acids (79.97 amu) were submitted for further confirmation and filtering. The corresponding MS/MS spectra of all the potential Ser/Thr phosphorylated peptides were examined manually for neutral loss of phosphoric acids. After eliminating the peptides that didn't show neutral loss (98 amu) in the MS/MS

spectra, further filter criteria was applied, including the following: i) No multiple post-translational modifications other than phosphorylation is allowed on the same peptide sequence; ii) Identified phosphorylated proteins must have at least one unmodified peptide assigned to them; iii) Xcorr score of phosphorylated peptides must be above 1.8 (single-charged), 2.2 (double-charged), or 2.5 (triple-charged). Because phosphorylated tyrosine doesn't generate neutral loss of phosphoric acids, the filter criteria for tyrosine phosphorylated peptides included all the above except for the observation of neutral loss in MS/MS spectra. After manual filtering, we identified 492 phosphorylation sites on 432 peptides corresponding to 288 different proteins (Table 2-1), of which 137 (47.08%) proteins were observed in one of the technical triplicates, 85 (29.21%) proteins in two, and 69 (23.71%) proteins in all three. Among the sequenced 432 phosphorylated peptides, there are 71.76% (310/432) present with multiple potential phosphorylation sites, and resulted in 52.44% (258/492) of the identified sites may not be confirmed due to the gas-phase rearrangements of phosphate groups during collision induced dissociation[18], although it has also been argued that the rearrangements should not affect phosphorylation site localization in any statistically significant way[19]. All of the identified phosphorylation sites were compared to the current literature using the Human Protein Reference Database (www.hprd.org) and UniProt Protein Knowledgebase (www.uniprot.org). Approximately 41% (200/492) of the assigned sites are not contained in the database and have been marked as novel assignments (Table 2-2).

Implicated kinases

In order to investigate which specific kinases are likely active in hESCs based on the phosphorylated proteome, the dataset was submitted to Scansite (scansite.mit.edu, 63

motifs available) for kinase prediction. With high-stringency scanning, we were able to discover 11 different motifs (4 motif families) interacting with 55 sites on 45 proteins from our dataset; with medium-stringency scanning, 18 different motifs (4 motif families) were predicted to be interacting with 172 sites on 109 proteins in the dataset (Table 2-3).

Functional classification

Approximately 86% (249/288) of the identified phosphorylated proteins were classified into 71 biological functions (21 networks) and 66 metabolic and signaling pathways. The top 5 assigned biological functions are: RNA post-transcriptional modification, cell cycle, cellular assembly and organization, DNA replication, recombination, and repair, and gene expression. While the top 5 involved canonical pathways are: apoptosis signaling, cell cycle G2/M DNA damage checkpoint regulation, nicotinate and nicotinamide metabolism, actin cytoskeleton signaling, and alanine metabolism.

Real-time electronic sensing of hESCs

To examine the impact of small molecules that affect the kinases CK2 or PKC, real time continual monitoring of hESCs was performed using a microsensor and impedance-based system. BG02 cells were plated and grown for 48 hours to establish subconfluent proliferating cultures. Compounds were added at 10 μM and cultures monitored for an additional 3 days. DMSO carrier control cultures continued to proliferate and were viable, as evidenced by increasing cell index after compound addition and stable cell index after 3 days in culture (Figure 2-5B and 2-5C, red trace). Small molecule inhibitors of CK2 (Figure 2-5A) slowed the increase in cell index slightly over the first 24 hours after addition, but their profiles were otherwise similar to the

carrier control (Figure 2-5B, green and blue traces). All three traces lay within the range of standard deviation (not shown), suggesting any impact of CK2 inhibition on hESCs was minimal. Conversely, addition of small molecule inhibitors, activators or modulators of PKC signaling (Figure 2-5A) had diverse and complex impact on hESCs. Two compounds, chelerythrine chloride and oleic acid, had no apparent effect on hESCs compared to the DMSO control (Figure 2-5C, blue, grey and red traces, respectively). 1-(5-Isoquinolinylsulfonyl)-3-methylpiperazine dihydrochloride caused a rise in impedance over the first 24 hours, while these cultures remained viable after 3 days (Figure 2-5C, black trace). Two compounds appeared to cause an arrest in the cell index, followed by general recovery on days 2-3 after addition: 1-(5-Isoquinolinylsulfonyl)-2-methylpiperazine dihydrochloride and Palmitoyl-DL-Carnitine chloride (Figure 2-5C, navy and pink traces, respectively). Four compounds, rac-2-Ethoxy-3-octadecanamido-1-propylphosphocholine, rac-2-Ethoxy-3-hexadecanamido-1-propylphosphocholine, Rottlerin and Phorbol 12-myristate 13-acetate, caused a general decrease in cell index over time, although with quite different profiles (Figure 2-5C, green, blue, dark green and brown traces, respectively). As overtly cytotoxic compounds cause a rapid drop in the cell index within 24-48 hours in this assay (not shown), the patterns caused by interference with PKC suggest specific affects on cellular characteristics. While the processes that were affected were not determined, it is likely that they could include proliferation, cell spreading, adhesion, calcium flux, apoptosis, or differentiation. It is notable that the affects of PKC interference were highly complex, for example, activators and inhibitors did not give characteristic outcomes. This suggests that in pluripotent cells

PKC signaling is highly complex, such that more than one PKC isotype may exert critical functions, or multiple downstream pathways may be regulated.

Assignment of the O-GlcNAc modified proteome within hESCs

Based on the dataset of identified proteins, all the peptides that had shown additional mass of one or more GlcNAcs (203.08 amu) were submitted for further filter and literature confirmation. The filter criteria applied for the O-GlcNAc proteome was described as above. Since O-GlcNAc modified peptides don't necessarily give out the neutral loss of GlcNAc, the dataset were not manually filtered by the observation of neutral loss in corresponding MS/MS spectra. After filtering, we identified 68 O-GlcNAc modification sites on 55 different proteins (Table 2-1), of which 45 (80.36%) were seen in one of the technical triplicate, 8 (14.29%) proteins were seen in two, and 3 (5.36%) proteins were seen in all three. All of the identified O-GlcNAc modified proteins were compared to the literature, and 18 of them have previously been identified as being O-GlcNAc modified by using various technologies while 37 proteins corresponding to 42 modification sites are considered to be novel (Table 2-4).

Functional classification

Over 98% (54/55) of the O-GlcNAc modified proteins were classified into 58 biological functions (6 networks) and 37 metabolic and signaling pathways. The top 5 assigned biological functions are: cellular assembly and organization, lipid metabolism, nucleic acid metabolism, small molecule biochemistry, and DNA replication, recombination, and repair and the top 5 involved canonical pathways are: valine, leucine and isoleucine biosynthesis, integrin signaling, alanine and aspartate metabolism, chemokine signaling and nucleotide sugar metabolism.

Assignment of hESCs secreted proteome

The secreted proteome of hESCs (BG02) was collected under defined media conditions. Tryptic digests of the secreted proteins were then separated by a 90-min SCX-LC experiment and combined into 5 fractions. 5 SCX fractions were then analyzed in 160-min capillary RP-LC coupled online to a linear ion trap-MS/MS-orbitrap instrument in triplicate for peptide assignment. Peptides and proteins were stringently filtered at 1% false discovery as described above. After eliminating duplicates and contaminants from cell-culture medium and basement membrane matrix (Matrigel™), the non-redundant combined dataset consists of 123 proteins identified by ~500 peptides (data not shown). By searching against HPRD, we were able to determine that 30 of the 123 proteins are supported by literature of being secreted. We suspected that the other 93 proteins might originate as intracellular proteins but being accidentally detected in the extracellular matrix caused by cell apoptosis. So we compared the 93 proteins with the 1000 most abundant proteins (ranked by spectral count) in our whole cell proteome dataset and were able to classify 57 of them into various subgroups of intracellular proteins (cytoplasm, nucleus, etc), while the rest of them remained undetermined. In addition, we had identified extracellular proteins during our whole cell proteomic analysis (2%, 74 proteins) and the combination of those 2 datasets resulted in a total of 102 secreted proteins from hESCs (Table 2-1).

Biological functions of identified proteins

96% (98/102) of the secreted proteins were classified into 73 biological functions (10 networks) and 68 metabolic and signaling pathways. The top 5 assigned biological functions are: cellular movement, cell death, genetic disorder, cellular assembly and

organization, and neurological disease. While the top 5 involved canonical pathways are: agrin interactions at neuromuscular junction, CDK5 signaling, acute phase response signaling, glycolysis/gluconeogenesis, and hypoxia signaling in cardiovascular system.

DISCUSSION

In this study, we combined different protein and peptide separation techniques with mass spectrometry to reduce the sample complexity and increase protein assignment and coverage of hESCs grown under defined media condition. This strategy resulted in the identification of 3189 proteins in hESCs including most of the well-known stem cell markers, such as Oct4 and Sox2. In our hESC sample, without intentional enrichment during experiments, we discovered 492 phosphorylation sites and 68 O-GlcNAc modification sites that have each been stringently filtered and manually examined. Furthermore, because of the unique feeder-free condition that the cells were cultured under, we were able to identify 102 proteins that are secreted from hESCs.

The proteome of hESCs

In order to validate our results and explore the scope of hESC proteome, we compared the identified proteins in our dataset to previous hESC proteomic studies that used similar techniques. Baharvand et al.[5] identified 685 proteins from 3 different hES cell lines (Royan H2, Royan H3 and Royan H5) by employing 2-DE combined with MALDI TOF-TOF mass spectrometry. By comparing with our dataset, only 75 (11%) proteins were found in common. Van Hoof et al.[4] identified 1775 proteins from undifferentiated hES cell line HES-2 by performing 2-DE separation and FT-ICR-MS/MS. After comparison, 1065 (60%) proteins were found in common with our dataset, which indicates certain consistency in the results both obtained from mass spectrometry-

based proteomic experiments. Finally, Schulz et al.[6] identified 589 proteins from undifferentiated hES cell line BG01 by using 934 antibodies in a large-scale western blotting system. By comparison, 134 (23%) proteins were found in common, and most of them were identified as relatively low-abundance transcription and translation regulators, receptors, and transporters, which can be explained by the specificity of antibodies used in the experiment. The discrepancy between datasets not only relates to the biological variations between cell lines or differences in analyzing techniques, but also reflects upon the complexity of hES cell proteome which has not been well-established and clearly needs deeper mining.

Several classes of proteins with particular relevance to hES cell biology were detected within the proteomic dataset. Nuclear factors that are known to regulate self-renewal, and others that mark the pluripotent state were identified including Oct4, Sox2, UTF1, DPPA4, and FoxD3 [20-23](Table 2-5).

The POU transcription factor Oct4 (encoded by POU5F1) plays a critical role in maintaining the undifferentiated state of ES cells[24-27]. It has been demonstrated that the level of Oct4 protein activity defines differentiation, de-differentiation, and self-renewal of ES cells, which suggests that Oct4 expression is regulated at the optimal level to ensure pluripotency in ES cells[25]. Sox2 is another transcription factor that is essential to the maintenance of self-renewal in pluripotent ES cells. Sox2 often works with Oct4 to regulate gene expression[28-29]. Recent studies showed that Sox2 is necessary for regulating multiple transcription factors that affect Oct4 expression, thus maintaining the requisite level of Oct4 for pluripotent state in ES cells[30]. Oct4 and Sox2 are also key participants in

reprogramming somatic cells back to undifferentiated states (iPSCs, induced pluripotent stem cells)[31-32].

Mammalian pluripotent cells including the zygote, blastomeres, inner cell mass, primitive ectoderm, primordial germ cells, gonadal germ cells and gametes form a developmental life cycle that ensures continued propagation of the species. Consequently, it is expected that these cells and their *in vitro* derivatives will have stringent mechanisms to ensure genomic integrity. While aneuploidies and genomic alterations can arise sporadically in hES cell cultures[33-34], hESCs have also been shown to be remarkably stable over prolonged culture periods[35]. We identified a network of DNA repair and replication proteins expressed by hESCs, consisting of DNA mismatch and double strand break repair factors, repair recruitment factors, chromosome cohesion and mitotic chromosome associated factors, as well as replication factors and s-phase checkpoint regulators. Our data lends mechanistic support to recent findings demonstrating that hESCs exhibit enhanced repair of multiple forms of DNA damage and resistance to ionizing radiation[36-37], highlighting systems that maintain genetic integrity in pluripotent cells.

Phosphorylated proteome of hESCs

Given the large number of phosphorylated peptides and novel phosporylation sites discovered, it was not surprising that several networks consisting entirely of phosphorylated proteins were assembled. β-Arrestin is an adaptor protein involved in the desensitization of G-protein coupled receptors, receptor internalization and recycling/degradation, as well as multiple other signaling functions in development and cellular movement[38]. Phosphorylation of β-arrestin at serine 412 is known to limit

activity to basal levels[39]. β-Arrestin (phospho-S412) was a focal point in a network of >30 phosphorylated proteins spanning the plasma membrane, cytoplasm and nucleus. Unmodified serine 412 peptides were not present, indicating that β-arrestin may be held in the basal state in undifferentiated cells. An alternate network of β-arrestin and non-phosphorylated proteins highlighted interactions with tubulins and staufen RNA localization factors. A network of phosphorylated proteins involved in chromatin epigenetics was also generated, including SWI/SNF chromatin regulators, members of the histone deacetylase complex, topoisomerases, DNA-interacting proteins, DNA-dependant protein kinase, splicing factors and a cohesin complex protein. These networks contain numerous newly identified phosphorylation sites and point to complex regulatory pathways that can be uncovered by deep proteomic mining.

The sequences of the known and novel phosphorylated peptides were interrogated for consensus kinase recognition/target sequences using Scansite (scansite.mit.edu, 63 motifs available). 11 different motifs, representing 4 motif families, were detected at 55 sites in 45 proteins using high stringency filtering. Of these, expression of the kinases CK2, GSK3β and a splice variant of CDK5 was detected in our dataset (Table 2-3), and others have been implicated as important for hESCs self-renewal or differentiation. The significance of AKT in hESCs self-renewal has been demonstrated in several ways. AKT signals downstream of PI3 kinase, and direct inhibition of either of these proteins causes differentiation of hESCs[40-41], as does inhibition of FRAP1/mTOR, which signals downstream of AKT[11]. Furthermore, we have shown that growth factors and receptors that activate this pathway, namely IGF/insulin and heregulin, support long term self-renewal of hESCs in defined media[11]. Similarly, a role for ERK1/2 signaling has been

suggested by MEK1/2-dependant activation which occurs following FGF2 stimulation[42] and increases in cell death and differentiation that are observed when MEK1/2 is inhibited[40]. Inhibition of Src-family kinases results in elevated differentiation of hESCs[43], and roles in self-renewal and differentiation of mouse ES cells have also been demonstrated[44]. Previously, activation of WNT signaling via inhibition of GSK3β was suggested to promote self-renewal in short term-assays[45]. However, subsequent reports have indicated that WNT/β-catenin signaling is not sufficient to maintain hESCs[46] and prolonged inhibition of GSK3β leads to mesendodermal differentiation[47]. Our analysis of protein phosphorylation within hESCs provides candidates that may be regulated by these kinases and a framework for continued dissection of the signaling pathways that control self-renewal and differentiation.

Detailed analyses of CK2 or PKC signaling in hESCs have not been reported. CK2 is a ubiquitous serine/threonine kinase involved with numerous cellular processes such as DNA replication, protein translation, the cell cycle and proliferation[48-49]. Mutation of the CK2 regulatory beta subunit leads to early embryonic lethality in the mouse, failure to expand ES cells from explanted blastocysts, or generation of null ES cells *in vitro*[50]. In contrast, inhibition of CK2 with two different small molecule antagonists had no overt effects on hESCs in our short-term assays. Fundamental differences between the signaling strategies of mouse and human pluripotent cells are suspected, for example hESCs are not dependant on the LIF/gp130 pathway[51]. The lack of obvious effects of CK2 inhibition on hESCs suggests a further point of species-specific signaling. PKC is a family of multifunctional serine/threonine kinases encompassing more than 10 different isoenzymes, with different cofactor and activation

requirements[52]. The detection of a PKC epsilon kinase recognition site with high stringency filtering, and additional PKC sites with medium stringency filtering, suggests functional activity for these kinases in hESCs or in differentiation. Increased proliferation of mouse ES cells in response to ATP has been suggested to involve PKC signaling[53], and S719 phosphorylation of PKC epsilon has been correlated with the proliferation rate of differentiating cells[54]. Proliferating hESCs with a range of small molecules that impact PKC signaling, either as inhibitors, modulators or activators, elicited complex response profiles. Therefore, it is possible that PKC family members have multiple important roles in hESCs, and knock-down of individual components in combination with agonist/antagonist probing will be required to unravel these potentially complex pathways.

O-GlcNAc modified proteins

Besides phosphorylation, we also took interest in glycosylation in hES cell proteome, specifically the O-GlcNAc modification. Glycosylation through serine and threonine by a single O-linked β-N-acetylglucosamine (O-GlcNAc) moiety is commonly seen among cytosolic and nuclear proteins in metazoans. O-GlcNAc modification is a nutrient/stress-sensitive modification that regulates proteins involved in a wide array of biological processes, including transcription, signal transduction, and metabolism[55-56]. It has been shown that aberrant regulation of O-GlcNAc levels is involved in cancer, insulin resistance, and several neurodegenerative diseases[57-58]. Furthermore, it has been demonstrated that a functional O-GlcNAc transferase (OGT), the enzyme that adds O-GlcNAc to serine or threonine residues, is essential for embryonic stem cell viability and mouse ontogeny[59]. O-GlcNAc modification shares several common traits with

phosphorylation, for example both of them modify serine and threonine residues, and have specialized enzymes for their addition to and removal from peptides (kinase and phosphatase for phosphate; O-GlcNAc transferase and O-GlcNAcase for O-GlcNAc). Studies have shown that O-GlcNAc modification exhibits complex and dynamic interactions with phosphorylation, such as competing for the same site, or being independent of the other[60]. In our dataset, we identified 68 O-GlcNAc modification sites on 55 proteins expanding a diverse range of functional classes including enzymes (UDP-glucose 6-dehydrogenase, fatty acid synthase), kinases (obscurin, TAO kinase), actin-based cytoskeletal proteins (talin), phosphatases (PPP1R12A), transcription regulators (general transcription factor II-I, elongation factor 2), transporters (Nup155, Nup160), and other undefined classes (histone, proliferating cell nuclear antigen). By comparing our results to the literature, which has established ~800 O-GlcNAc modified proteins with less than 12% site-mapping on them over the past 30 years[61-65], 18 proteins in our dataset were confirmed as O-GlcNAc modified proteins by other studies and 37 proteins were identified as novel O-GlcNAc modified proteins. Combined from both confirmed and novel O-GlcNAc proteins, a total of 68 novel O-GlcNAc modification sites were revealed. We also discovered that 16 proteins are both modified by O-phosphate and O-GlcNAc (Table 2-6), including lamin A/C, and nuclear pore complex proteins such as Nup358 and Nup160, etc. Lamins are structural protein components of the nuclear lamina, a protein network underlying the inner nuclear membrane that determines nuclear shape and size. It has been proven that lamina modulates the extent of deformation of nucleus and therefore makes nuclei in undifferentiated stem cells physically plastic and more pliable than nuclei in differentiated cells[66]. Furthermore, it has been shown that the

expression of lamin A/C is activated during human ES cell differentiation before the down-regulation of the pluripotency marker Oct4 but not before the down-regulation of the pluripotency markers Tra-1-60, Tra-1-81, and SSEA-4. Those findings identified the absence of lamin A/C expression as a novel marker for undifferentiated ES cells and further support a role for nuclear lamins in cell maintenance and differentiation[67]. Besides nuclear lamins, nuclear pore complexes, which are responsible for the protected exchange of components between the nucleus and cytoplasm and for preventing the transport of material not destined to cross the nuclear envelope, have also been demonstrated to be involved in stem cell cardiovascular differentiation. Nuclear pore complex proteins exhibits a significantly decreased density and depth in stem cell-derived cardiomyocytes, indicating that stem cell-derived cardiomyocytes undergo structural adaptation and mobilize nuclear transport regulators in support of nucleocytoplasmic communication during commitment to mature cardiac lineage[68].

Comparison of the results between multidimensional protein identification technology (MudPIT) and single dimension LC-MS/MS experiment (1% FDR)

Post-translational modification of proteins, such as phosphorylation and glycosylation, often plays important roles in regulation of biological processes in eukaryotes. Detecting and identifying post-translationally modified proteins facilitates the elucidation of the mechanisms in cellular signal transduction and other biological activities. Despite the recent advancement of mass spectrometry-based proteomics, identifying post-translationally modified proteins, especially characterizing the modification sites, still remains a challenge since most post-translationally modified proteins are naturally expressed at low abundance. In order to increase the coverage of

post-translationally modified proteins in the context of a complex cell proteome, certain enrichment techniques, such as chromatography, become necessary prior to protein/peptide sequencing by mass spectrometer. In our study, we compared the data of identified cellular proteome, particularly the post-translationally modified proteome, obtained respectively from a single LC-MS/MS experiment and the MudPIT experiments described in the previous method section. By comparing the numbers of identified proteins in two respective analyses, we observed an overall 2.26 fold increase (2414/1067) in identified proteins with a 3.7 fold increase (21523/5821) in total peptides (Figure 2-6A) in the MudPIT experiments. Particularly, both of phosphate and O-GlcNAc modified proteomes showed a drastic increase for over an order of magnitude in identified proteins and corresponding peptides in the MudPIT compared to single LC-MS/MS experiments (Figure 2-6B). For instance, the two proteins that are identified as being modified by both phosphate and O-GlcNAc, fatty acid synthase (FASN) and nucleolin (NCL), both showed approximately 2-fold increase in the numbers of unique peptides assigned to them (Table 2-S1). Furthermore, 94.4% (1007/1067) of identified proteins in the single LC-MS/MS experiment were also present in the MudPIT analyses, and 65.7% (701/1067) of them were observed in the 1000 most abundant proteins in the MudPIT experiment (ranked by spectral count). Those data indicates that multidimensional separation of samples prior to MS/MS sequencing not only increases the number of identified proteins but also the peptide coverage of identified proteins leading to more confident protein identification. The significant increase in modified protein identification also demonstrates that this technique improves the depth of

proteomic analysis of complex mixture by enriching low-abundance, post-translationally modified proteins.

Secreted proteome of hESCs

Previously, the extracellular matrix (ECM) proteins that mediate hES cell adherence and matrix-integrin signaling have been inferred by the combination of integrins expressed by hESCs and plating experiments with purified proteins. Laminin and fibronectin have therefore been used to support serial attachment and expansion of hESCs[69-70]. Our analysis of the secreted proteome identified a diverse array of human ECM proteins: tenascin X, fibulin, Laminin alpha 2 and alpha 5, and collagen alpha I, IV, and XVIII (Table 2-7). It is likely that hESCs deposit their own ECM after attachment and combinations of these proteins may provide more effective substrates for adhesion.

Transcription factor LIN28 was identified in our ECM dataset (Table 2-7), which modulates cell growth and associates with a subset of cell cycle regulators in embryonic stem cells and therefore can be used as markers for cell pluripotency[71-72]. We also observed several growth factors that are important in signal transduction in ESCs, such as Jagged-1, GDF3, and LEFTY (Table 2-7). Jagged-1 is a Notch ligand which functions as WNT-dependent Notch signaling activator in ESCs and is the key molecule maintaining the homeostasis of the cells[73]. In previous studies, three secreted factors of the TGFβ family, LEFTY1, LEFTY2 and GDF3, are expressed at high levels during pluripotency and decline dramatically when the cells are allowed to differentiate, suggesting a role during this process[74]. It was further proven that GDF3 regulates both of the two major characteristics of embryonic stem cells in a species-specific manner: the ability to maintain the undifferentiated state and the ability to differentiate into the full spectrum of

cell types[75]. LEFTY was shown to be regulated by both pluripotency-supporting pathways (Smad, WNT) and transcription factor Oct4, and also induced upon the exit of pluripotent state[76-79]. Lastly, the observation of hESCs forming polarized epithelia[6, 80] was confirmed with the detection of a substantial number of tight-junction signaling complex proteins in our dataset.

SUMMARY

These experiments serve to investigate the proteome of hESCs in order to discover proteomic candidates that are responsible for the maintenance of undifferentiated state of the cells and the initial loss of pluripotency. By performing a multidimensional chromatography separation coupled with tandem mass spectrometry-based protein identification experiment, we were able to assign the cell proteome with a certain depth and specific post-translational modifications within the cell that are significant to various cellular processes. To further reveal the molecular mechanism and signal transduction pathways involved in hESC differentiation, proteomic variations between undifferentiated cells and hESC-derived specialized cell types will need to be elucidated. Our dataset of the hES cell proteome was analyzed in triplicate and clustered with spectral counts, which will enable us to perform non-isotope-based relative quantification in a statistically significant manner to compare the proteome and post-translational modifications of these hESCs to any cell lineage derived from these pluripotent cells.

REFERENCES

1. Metallo, C. M.; Azarin, S. M.; Ji, L.; de Pablo, J. J.; Palecek, S. P., Engineering tissue from human embryonic stem cells. *J Cell Mol Med* **2008,** 12, (3), 709-29.

2. Abuljadayel, I. S., Harnessing pluripotency from differentiated cells: a regenerative source for tissue-specific stem cell therapies. *Curr Stem Cell Res Ther* **2006,** 1, (3), 325-31.

3. Srivastava, D.; Ivey, K. N., Potential of stem-cell-based therapies for heart disease. *Nature* **2006,** 441, (7097), 1097-9.

4. Van Hoof, D.; Passier, R.; Ward-Van Oostwaard, D.; Pinkse, M. W.; Heck, A. J.; Mummery, C. L.; Krijgsveld, J., A quest for human and mouse embryonic stem cell-specific proteins. *Mol Cell Proteomics* **2006,** 5, (7), 1261-73.

5. Baharvand, H.; Hajheidari, M.; Ashtiani, S. K.; Salekdeh, G. H., Proteomic signature of human embryonic stem cells. *Proteomics* **2006,** 6, (12), 3544-9.

6. Schulz, T. C.; Swistowska, A. M.; Liu, Y.; Swistowski, A.; Palmarini, G.; Brimble, S. N.; Sherrer, E.; Robins, A. J.; Rao, M. S.; Zeng, X., A large-scale proteomic analysis of human embryonic stem cells. *BMC Genomics* **2007,** 8, 478.

7. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd, Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **1999,** 17, (7), 676-82.

8. Moore, A. W., Jr.; Larmann, J. P., Jr.; Lemmo, A. V.; Jorgenson, J. W., Two-dimensional liquid chromatography-capillary electrophoresis techniques for analysis of proteins and peptides. *Methods Enzymol* **1996,** 270, 401-19.

9. Opiteck, G. J.; Jorgenson, J. W., Two-dimensional SEC/RPLC coupled to mass spectrometry for the analysis of peptides. *Anal Chem* **1997,** 69, (13), 2283-91.

10. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **2001,** 19, (3), 242-7.

11. Wang, L.; Schulz, T. C.; Sherrer, E. S.; Dauphin, D. S.; Shin, S.; Nelson, A. M.; Ware, C. B.; Zhan, M.; Song, C. Z.; Chen, X.; Brimble, S. N.; McLean, A.; Galeano, M. J.; Uhl, E. W.; D'Amour, K. A.; Chesnut, J. D.; Rao, M. S.; Blau, C. A.; Robins, A. J., Self-renewal of human embryonic stem cells requires insulin-like growth factor-1 receptor and ERBB2 receptor signaling. *Blood* **2007,** 110, (12), 4111-9.

12. Robins, A. J., Schulz, T. C., Novel methods of Stem Cell Culture and Maintenance: Media and extra cellular matrix requirements for large scale ESC growth. In *Emerging Technology Platforms for Stem Cells*, John Wiley & Sons: 2008.

13. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004,** 76, (14), 4193-201.

14. Weatherly, D. B.; Atwood, J. A., 3rd; Minning, T. A.; Cavola, C.; Tarleton, R. L.; Orlando, R., A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics* **2005,** 4, (6), 762-72.

15. Ganter, B.; Giroux, C. N., Emerging applications of network and pathway analysis in drug discovery and development. *Curr Opin Drug Discov Devel* **2008,** 11, (1), 86-94.

16. Ganter, B.; Zidek, N.; Hewitt, P. R.; Muller, D.; Vladimirova, A., Pathway analysis tools and toxicogenomics reference databases for risk assessment. *Pharmacogenomics* **2008,** 9, (1), 35-54.

17. Solly, K.; Wang, X.; Xu, X.; Strulovici, B.; Zheng, W., Application of real-time cell electronic sensing (RT-CES) technology to cell-based assays. *Assay Drug Dev Technol* **2004,** 2, (4), 363-72.

18. Palumbo, A. M.; Reid, G. E., Evaluation of Gas-Phase Rearrangement and Competing Fragmentation Reactions on Protein Phosphorylation Site Assignment Using Collision Induced Dissociation-MS/MS and MS(3). *Anal Chem* **2008**.

19. Aguiar, M.; Haas, W.; Beausoleil, S. A.; Rush, J.; Gygi, S. P., Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets. *J Proteome Res* **2010,** 9, (6), 3103-7.

20. van den Boom, V.; Kooistra, S. M.; Boesjes, M.; Geverts, B.; Houtsmuller, A. B.; Monzen, K.; Komuro, I.; Essers, J.; Drenth-Diephuis, L. J.; Eggen, B. J., UTF1 is a chromatin-associated protein involved in ES cell differentiation. *J Cell Biol* **2007,** 178, (6), 913-24.

21. Liu, Y.; Labosky, P. A., Regulation of embryonic stem cell self-renewal and pluripotency by Foxd3. *Stem Cells* **2008,** 26, (10), 2475-84.

22. Boyer, L. A.; Lee, T. I.; Cole, M. F.; Johnstone, S. E.; Levine, S. S.; Zucker, J. P.; Guenther, M. G.; Kumar, R. M.; Murray, H. L.; Jenner, R. G.; Gifford, D. K.; Melton, D.

A.; Jaenisch, R.; Young, R. A., Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **2005,** 122, (6), 947-56.

23. Babaie, Y.; Herwig, R.; Greber, B.; Brink, T. C.; Wruck, W.; Groth, D.; Lehrach, H.; Burdon, T.; Adjaye, J., Analysis of Oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. *Stem Cells* **2007,** 25, (2), 500-10.

24. Chambers, I.; Colby, D.; Robertson, M.; Nichols, J.; Lee, S.; Tweedie, S.; Smith, A., Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **2003,** 113, (5), 643-55.

25. Nichols, J.; Zevnik, B.; Anastassiadis, K.; Niwa, H.; Klewe-Nebenius, D.; Chambers, I.; Scholer, H.; Smith, A., Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **1998,** 95, (3), 379-91.

26. Niwa, H.; Miyazaki, J.; Smith, A. G., Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* **2000,** 24, (4), 372-6.

27. Mitsui, K.; Tokuzawa, Y.; Itoh, H.; Segawa, K.; Murakami, M.; Takahashi, K.; Maruyama, M.; Maeda, M.; Yamanaka, S., The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **2003,** 113, (5), 631-42.

28. Rodda, D. J.; Chew, J. L.; Lim, L. H.; Loh, Y. H.; Wang, B.; Ng, H. H.; Robson, P., Transcriptional regulation of nanog by OCT4 and SOX2. *J Biol Chem* **2005,** 280, (26), 24731-7.

29. Ambrosetti, D. C.; Basilico, C.; Dailey, L., Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol Cell Biol* **1997,** 17, (11), 6321-9.

30. Masui, S.; Nakatake, Y.; Toyooka, Y.; Shimosato, D.; Yagi, R.; Takahashi, K.; Okochi, H.; Okuda, A.; Matoba, R.; Sharov, A. A.; Ko, M. S.; Niwa, H., Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol* **2007,** 9, (6), 625-35.

31. Takahashi, K.; Yamanaka, S., Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **2006,** 126, (4), 663-76.

32. Yu, J.; Vodyanik, M. A.; Smuga-Otto, K.; Antosiewicz-Bourget, J.; Frane, J. L.; Tian, S.; Nie, J.; Jonsdottir, G. A.; Ruotti, V.; Stewart, R.; Slukvin, II; Thomson, J. A., Induced pluripotent stem cell lines derived from human somatic cells. *Science* **2007,** 318, (5858), 1917-20.

33. Draper, J. S.; Smith, K.; Gokhale, P.; Moore, H. D.; Maltby, E.; Johnson, J.; Meisner, L.; Zwaka, T. P.; Thomson, J. A.; Andrews, P. W., Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nat Biotechnol* **2004,** 22, (1), 53-4.

34. Maitra, A.; Arking, D. E.; Shivapurkar, N.; Ikeda, M.; Stastny, V.; Kassauei, K.; Sui, G.; Cutler, D. J.; Liu, Y.; Brimble, S. N.; Noaksson, K.; Hyllner, J.; Schulz, T. C.; Zeng, X.; Freed, W. J.; Crook, J.; Abraham, S.; Colman, A.; Sartipy, P.; Matsui, S.; Carpenter, M.; Gazdar, A. F.; Rao, M.; Chakravarti, A., Genomic alterations in cultured human embryonic stem cells. *Nat Genet* **2005,** 37, (10), 1099-103.

35. Buzzard, J. J.; Gough, N. M.; Crook, J. M.; Colman, A., Karyotype of human ES cells during extended culture. *Nat Biotechnol* **2004,** 22, (4), 381-2; author reply 382.

36. Maynard, S.; Swistikowa, A. M.; Lee, J. W.; Liu, Y.; Liu, S. T.; A, D. A. C.; Rao, M.; de Souza-Pinto, N.; Zeng, X.; Bohr, V. A., Human Embryonic Stem Cells have Enhanced Repair of Multiple Forms of DNA Damage. *Stem Cells* **2008**.

37. Banuelos, C. A.; Banath, J. P.; Macphail, S. H.; Zhao, J.; Eaves, C. A.; O'Connor, M. D.; Lansdorp, P. M.; Olive, P. L., Mouse but not human embryonic stem cells are deficient in rejoining of ionizing radiation-induced DNA double-strand breaks. *DNA Repair (Amst)* **2008**.

38. Buchanan, F. G.; DuBois, R. N., Emerging roles of beta-arrestins. *Cell Cycle* **2006,** 5, (18), 2060-3.

39. Luttrell, L. M.; Ferguson, S. S.; Daaka, Y.; Miller, W. E.; Maudsley, S.; Della Rocca, G. J.; Lin, F.; Kawakatsu, H.; Owada, K.; Luttrell, D. K.; Caron, M. G.; Lefkowitz, R. J., Beta-arrestin-dependent formation of beta2 adrenergic receptor-Src protein kinase complexes. *Science* **1999,** 283, (5402), 655-61.

40. Armstrong, L.; Hughes, O.; Yung, S.; Hyslop, L.; Stewart, R.; Wappler, I.; Peters, H.; Walter, T.; Stojkovic, P.; Evans, J.; Stojkovic, M.; Lako, M., The role of PI3K/AKT, MAPK/ERK and NFkappabeta signalling in the maintenance of human embryonic stem cell pluripotency and viability highlighted by transcriptional profiling and functional analysis. *Hum Mol Genet* **2006,** 15, (11), 1894-913.

41. McLean, A. B.; D'Amour, K. A.; Jones, K. L.; Krishnamoorthy, M.; Kulik, M. J.; Reynolds, D. M.; Sheppard, A. M.; Liu, H.; Xu, Y.; Baetge, E. E.; Dalton, S., Activin a

efficiently specifies definitive endoderm from human embryonic stem cells only when phosphatidylinositol 3-kinase signaling is suppressed. *Stem Cells* **2007,** 25, (1), 29-38.

42. Kang, H. B.; Kim, J. S.; Kwon, H. J.; Nam, K. H.; Youn, H. S.; Sok, D. E.; Lee, Y., Basic fibroblast growth factor activates ERK and induces c-fos in human embryonic stem cell line MizhES1. *Stem Cells Dev* **2005,** 14, (4), 395-401.

43. Anneren, C.; Cowan, C. A.; Melton, D. A., The Src family of tyrosine kinases is important for embryonic stem cell self-renewal. *J Biol Chem* **2004,** 279, (30), 31590-8.

44. Meyn, M. A., 3rd; Schreiner, S. J.; Dumitrescu, T. P.; Nau, G. J.; Smithgall, T. E., SRC family kinase activity is required for murine embryonic stem cell growth and differentiation. *Mol Pharmacol* **2005,** 68, (5), 1320-30.

45. Sato, N.; Meijer, L.; Skaltsounis, L.; Greengard, P.; Brivanlou, A. H., Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat Med* **2004,** 10, (1), 55-63.

46. Dravid, G.; Ye, Z.; Hammond, H.; Chen, G.; Pyle, A.; Donovan, P.; Yu, X.; Cheng, L., Defining the role of Wnt/beta-catenin signaling in the survival, proliferation, and self-renewal of human embryonic stem cells. *Stem Cells* **2005,** 23, (10), 1489-501.

47. Bakre, M. M.; Hoi, A.; Mong, J. C.; Koh, Y. Y.; Wong, K. Y.; Stanton, L. W., Generation of multipotential mesendodermal progenitors from mouse embryonic stem cells via sustained Wnt pathway activation. *J Biol Chem* **2007,** 282, (43), 31703-12.

48. Pinna, L. A.; Meggio, F., Protein kinase CK2 ("casein kinase-2") and its implication in cell division and proliferation. *Prog Cell Cycle Res* **1997,** 3, 77-97.

49. Guerra, B.; Issinger, O. G., Protein kinase CK2 and its role in cellular proliferation, development and pathology. *Electrophoresis* **1999,** 20, (2), 391-408.

50. Buchou, T.; Vernet, M.; Blond, O.; Jensen, H. H.; Pointu, H.; Olsen, B. B.; Cochet, C.; Issinger, O. G.; Boldyreff, B., Disruption of the regulatory beta subunit of protein kinase CK2 in mice leads to a cell-autonomous defect and early embryonic lethality. *Mol Cell Biol* **2003,** 23, (3), 908-15.

51. Okita, K.; Yamanaka, S., Intracellular signaling pathways regulating pluripotency of embryonic stem cells. *Curr Stem Cell Res Ther* **2006,** 1, (1), 103-11.

52. Breitkreutz, D.; Braiman-Wiksman, L.; Daum, N.; Denning, M. F.; Tennenbaum, T., Protein kinase C family: on the crossroads of cell signaling in skin and tumor epithelium. *J Cancer Res Clin Oncol* **2007,** 133, (11), 793-808.

53. Heo, J. S.; Han, H. J., ATP stimulates mouse embryonic stem cell proliferation via protein kinase C, phosphatidylinositol 3-kinase/Akt, and mitogen-activated protein kinase signaling pathways. *Stem Cells* **2006,** 24, (12), 2637-48.

54. Prudhomme, W.; Daley, G. Q.; Zandstra, P.; Lauffenburger, D. A., Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. *Proc Natl Acad Sci U S A* **2004,** 101, (9), 2900-5.

55. Love, D. C.; Hanover, J. A., The hexosamine signaling pathway: deciphering the "O-GlcNAc code". *Sci STKE* **2005,** 2005, (312), re13.

56. Zachara, N. E.; Hart, G. W., O-GlcNAc a sensor of cellular state: the role of nucleocytoplasmic glycosylation in modulating cellular function in response to nutrition and stress. *Biochim Biophys Acta* **2004,** 1673, (1-2), 13-28.

57. Wells, L.; Vosseller, K.; Hart, G. W., Glycosylation of nucleocytoplasmic proteins: signal transduction and O-GlcNAc. *Science* **2001,** 291, (5512), 2376-8.

58. Zachara, N. E.; Hart, G. W., The emerging significance of O-GlcNAc in cellular regulation. *Chem Rev* **2002,** 102, (2), 431-8.

59. Shafi, R.; Iyer, S. P.; Ellies, L. G.; O'Donnell, N.; Marek, K. W.; Chui, D.; Hart, G. W.; Marth, J. D., The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc Natl Acad Sci U S A* **2000,** 97, (11), 5735-9.

60. Slawson, C.; Hart, G. W., Dynamic interplay between O-GlcNAc and O-phosphate: the sweet side of protein regulation. *Curr Opin Struct Biol* **2003,** 13, (5), 631-6.

61. Copeland, R. J.; Bullen, J. W.; Hart, G. W., Cross-talk between GlcNAcylation and phosphorylation: roles in insulin resistance and glucose toxicity. *Am J Physiol Endocrinol Metab* **2008,** 295, (1), E17-28.

62. Khidekel, N.; Ficarro, S. B.; Peters, E. C.; Hsieh-Wilson, L. C., Exploring the O-GlcNAc proteome: direct identification of O-GlcNAc-modified proteins from the brain. *Proc Natl Acad Sci U S A* **2004,** 101, (36), 13132-7.

63. Teo, C. F.; Ingale, S.; Wolfert, M. A.; Elsayed, G. A.; Not, L. G.; Chatham, J. C.; Wells, L.; Boons, G. J., Glycopeptide-specific monoclonal antibodies suggest new roles for O-GlcNAc. *Nat Chem Biol* **2010**.

64. Wang, Z.; Udeshi, N. D.; Slawson, C.; Compton, P. D.; Sakabe, K.; Cheung, W. D.; Shabanowitz, J.; Hunt, D. F.; Hart, G. W., Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Sci Signal* **2010,** 3, (104), ra2.

65. Whelan, S. A.; Hart, G. W., Proteomic approaches to analyze the dynamic relationships between nucleocytoplasmic protein glycosylation and phosphorylation. *Circ Res* **2003,** 93, (11), 1047-58.

66. Pajerowski, J. D.; Dahl, K. N.; Zhong, F. L.; Sammak, P. J.; Discher, D. E., Physical plasticity of the nucleus in stem cell differentiation. *Proc Natl Acad Sci U S A* **2007,** 104, (40), 15619-24.

67. Constantinescu, D.; Gray, H. L.; Sammak, P. J.; Schatten, G. P.; Csoka, A. B., Lamin A/C expression is a marker of mouse and human embryonic stem cell differentiation. *Stem Cells* **2006,** 24, (1), 177-85.

68. Perez-Terzic, C.; Behfar, A.; Mery, A.; van Deursen, J. M.; Terzic, A.; Puceat, M., Structural adaptation of the nuclear pore complex in stem cell-derived cardiomyocytes. *Circ Res* **2003,** 92, (4), 444-52.

69. Xu, C.; Inokuma, M. S.; Denham, J.; Golds, K.; Kundu, P.; Gold, J. D.; Carpenter, M. K., Feeder-free growth of undifferentiated human embryonic stem cells. *Nat Biotechnol* **2001,** 19, (10), 971-4.

70. Brimble, S. N.; Zeng, X.; Weiler, D. A.; Luo, Y.; Liu, Y.; Lyons, I. G.; Freed, W. J.; Robins, A. J.; Rao, M. S.; Schulz, T. C., Karyotypic stability, genotyping, differentiation, feeder-free maintenance, and gene expression sampling in three human embryonic stem cell lines derived prior to August 9, 2001. *Stem Cells Dev* **2004,** 13, (6), 585-97.

71. Richards, M.; Tan, S. P.; Tan, J. H.; Chan, W. K.; Bongso, A., The transcriptome profile of human embryonic stem cells as defined by SAGE. *Stem Cells* **2004,** 22, (1), 51-64.

72. Assou, S.; Cerecedo, D.; Tondeur, S.; Pantesco, V.; Hovatta, O.; Klein, B.; Hamamah, S.; De Vos, J., A gene expression signature shared by human mature oocytes and embryonic stem cells. *BMC Genomics* **2009,** 10, 10.

73. Katoh, M., Notch ligand, JAG1, is evolutionarily conserved target of canonical WNT signaling pathway in progenitor cells. *Int J Mol Med* **2006,** 17, (4), 681-5.

74. Sato, N.; Sanjuan, I. M.; Heke, M.; Uchida, M.; Naef, F.; Brivanlou, A. H., Molecular signature of human embryonic stem cells and its comparison with the mouse. *Dev Biol* **2003,** 260, (2), 404-13.

75. Levine, A. J.; Brivanlou, A. H., GDF3, a BMP inhibitor, regulates cell fate in stem cells and early embryos. *Development* **2006,** 133, (2), 209-16.

76. Katoh, M., WNT/PCP signaling pathway and human cancer (review). *Oncol Rep* **2005,** 14, (6), 1583-8.

77. Meijer, L.; Skaltsounis, A. L.; Magiatis, P.; Polychronopoulos, P.; Knockaert, M.; Leost, M.; Ryan, X. P.; Vonica, C. A.; Brivanlou, A.; Dajani, R.; Crovace, C.; Tarricone, C.; Musacchio, A.; Roe, S. M.; Pearl, L.; Greengard, P., GSK-3-selective inhibitors derived from Tyrian purple indirubins. *Chem Biol* **2003,** 10, (12), 1255-66.

78. Cserjesi, P.; Brown, D.; Lyons, G. E.; Olson, E. N., Expression of the novel basic helix-loop-helix gene eHAND in neural crest derivatives and extraembryonic membranes during mouse development. *Dev Biol* **1995,** 170, (2), 664-78.

79. Beck, F.; Erler, T.; Russell, A.; James, R., Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. *Dev Dyn* **1995,** 204, (3), 219-27.

80. Krtolica, A.; Genbacev, O.; Escobedo, C.; Zdravkovic, T.; Nordstrom, A.; Vabuena, D.; Nath, A.; Simon, C.; Mostov, K.; Fisher, S. J., Disruption of apical-basal polarity of human embryonic stem cells enhances hematoendothelial differentiation. *Stem Cells* **2007,** 25, (9), 2215-23.

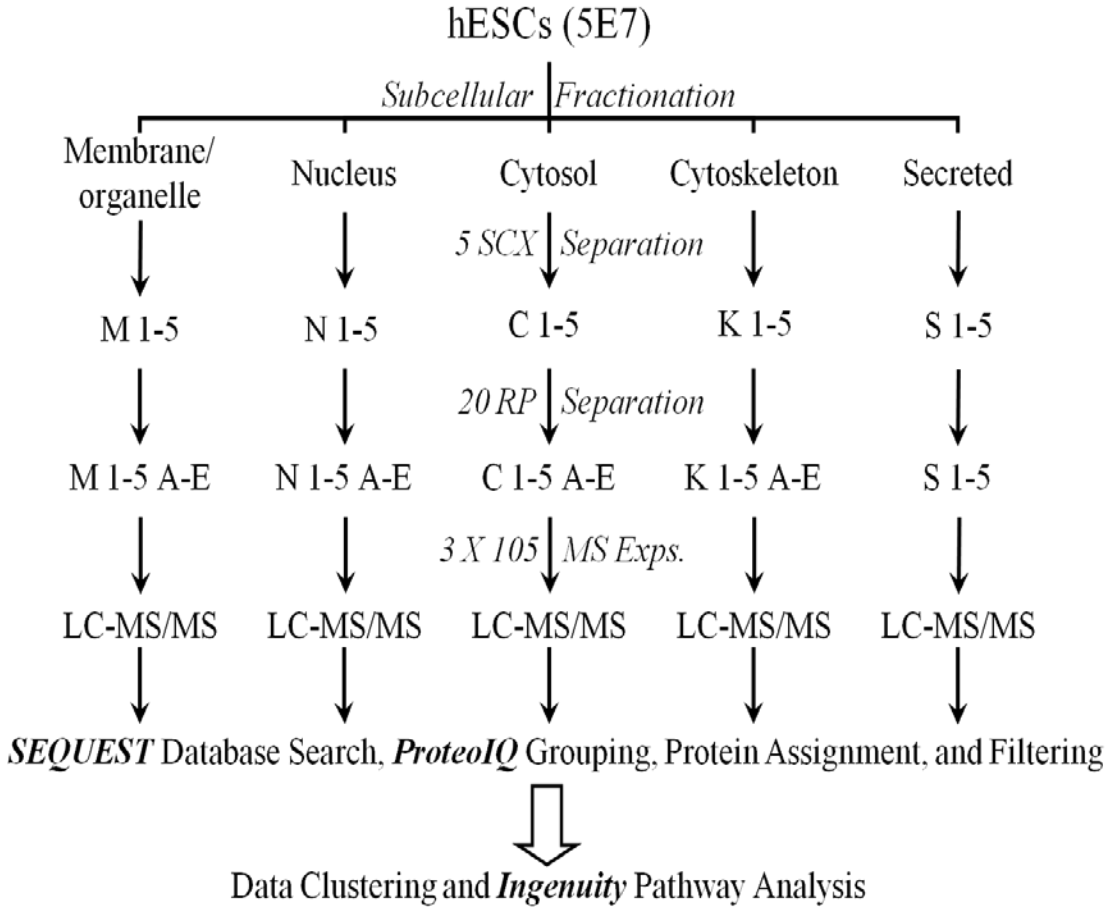
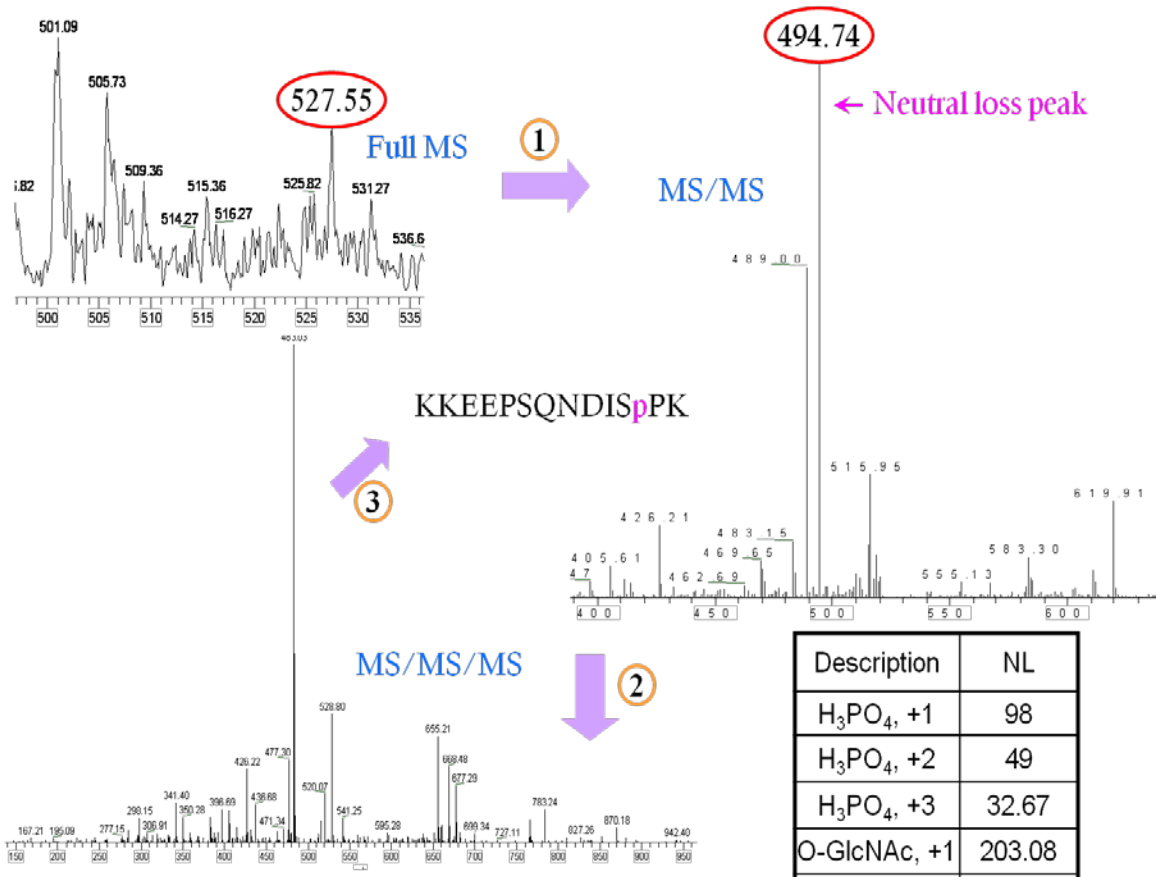


Figure 2-1. Experiment workflow.
Proteins extracted from hESCs were analyzed with SCX-LC, RP-LC and nanoLC-MS/MS to establish a 3-dimension separation prior to MS/MS peptide sequencing.



Description	NL
H_3PO_4, +1	98
H_3PO_4, +2	49
H_3PO_4, +3	32.67
O-GlcNAc, +1	203.08
O-GlcNAc, +2	101.55
O-GlcNAc, +3	67.67

Figure 2-2. Neutral loss-dependent MS^3 method.
The detection of any neutral loss listed in the table between the precursor ion in full MS and the 3 most intense product ions in MS^2 automatically triggers the subsequent MS^3 fragmentation.

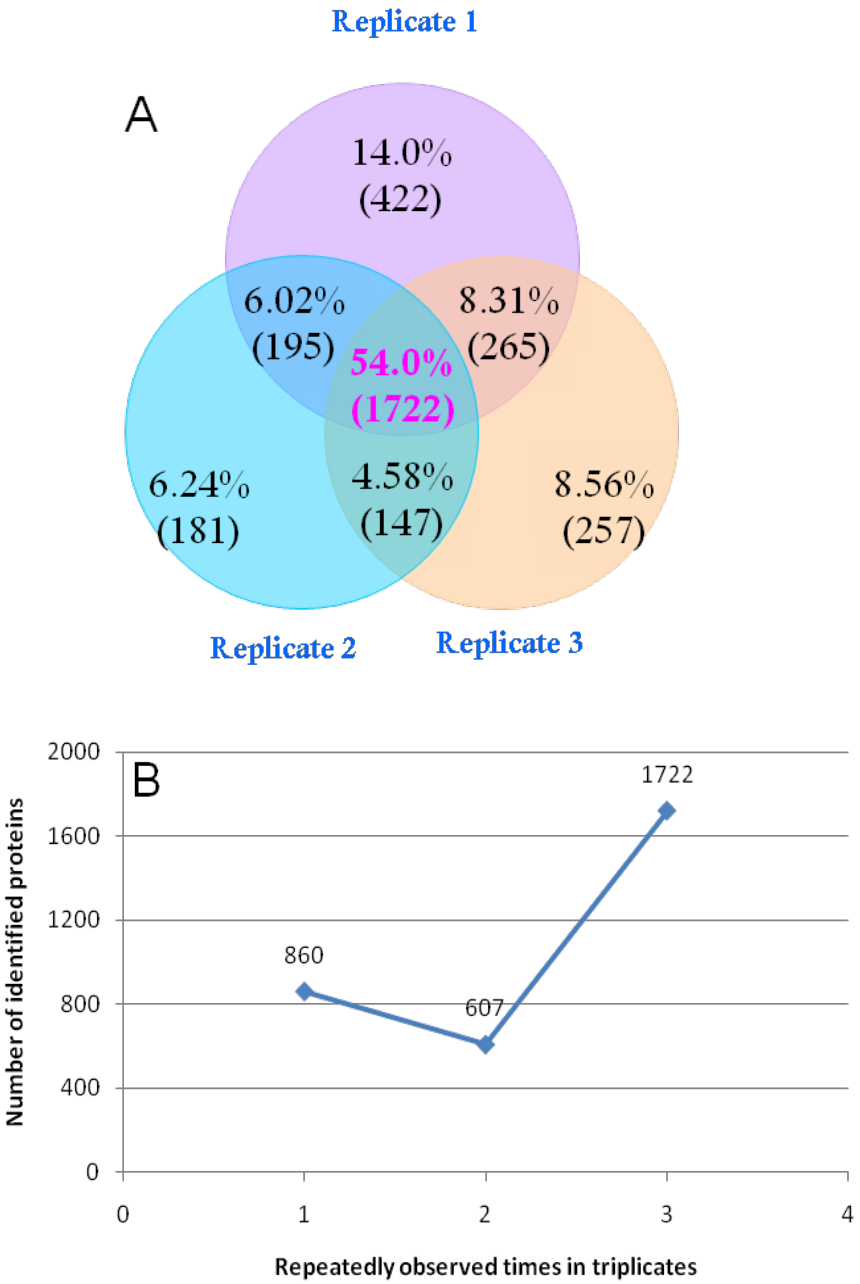


Figure 2-3. Statistics across LC-MS/MS technically triplicate experiments.
(**A**) Reproducibility between technical replicates; (**B**) Cumulative number of identified proteins plotted against the number of LC-MS/MS replicate experiments. 1722 proteins (54%) were identified in each of the triplicate, 860 proteins (422, 181, 257 proteins, respectively) were observed in one of the triplicate, and 607 proteins (195, 147, 265 proteins, respectively) were observed in two of the triplicate.

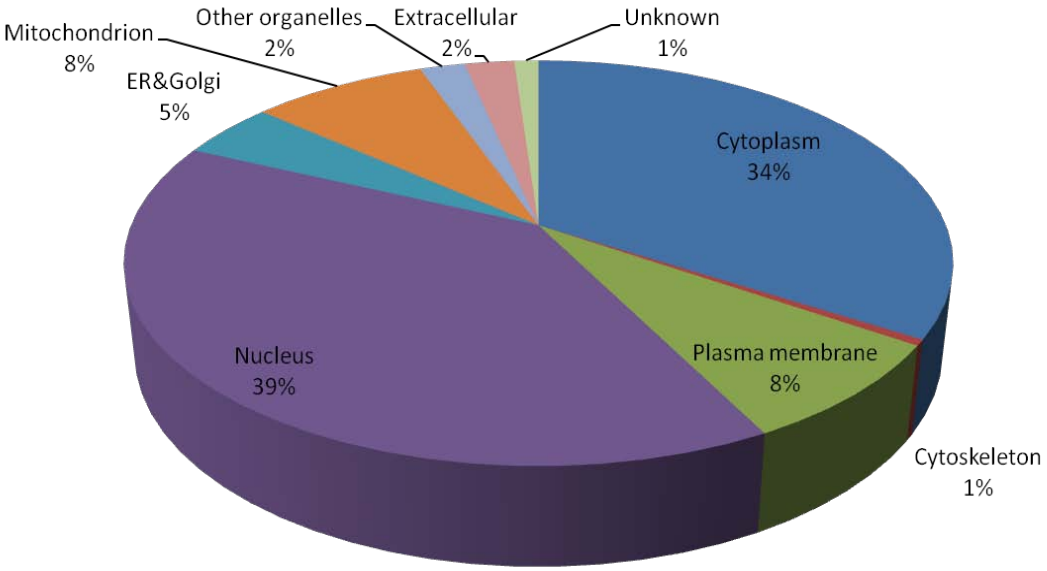


Figure 2-4. Subcellular localization of proteins identified from hESCs.
The majority of identified proteins were classified as nuclear and cytoplasmic proteins.

A

Compound	Class	Action	Sigma Cat.	Panel	Color
DMSO				B/C	Red
CK2 Inhibitor 2	CK2	Inhibitor	C 7367	B	Green
TBBz	CK2	Inhibitor	T 6951	B	Blue
rac-2-Ethoxy-3-hexadecanamido-1-propylphosphocholine	PKC	Inhibitor	E 3645	C	Blue
rac-2-Ethoxy-3-octadecanamido-1-propylphosphocholine	PKC	Inhibitor	E 3770	C	Green
1-(5-Isoquinolinylsulfonyl)-3-methylpiperazine dihydrochloride	PKC	Inhibitor	I 6391	C	Black
1-(5-Isoquinolinylsulfonyl)-2-methylpiperazine dihydrochloride	PKA / PKC	Inhibitor	I 7016	C	Navy
Rottlerin	PKC / CaM	Inhibitor	R 5648	C	Dark Green
Oleic Acid	PKC	Activator	O 1008	C	Grey
Phorbol 12-myristate 13-acetate	PKC	Activator	P 8139	C	Brown
Palmitoyl-DL-Carnitine chloride	PKC	Modulator	P 4509	C	Pink

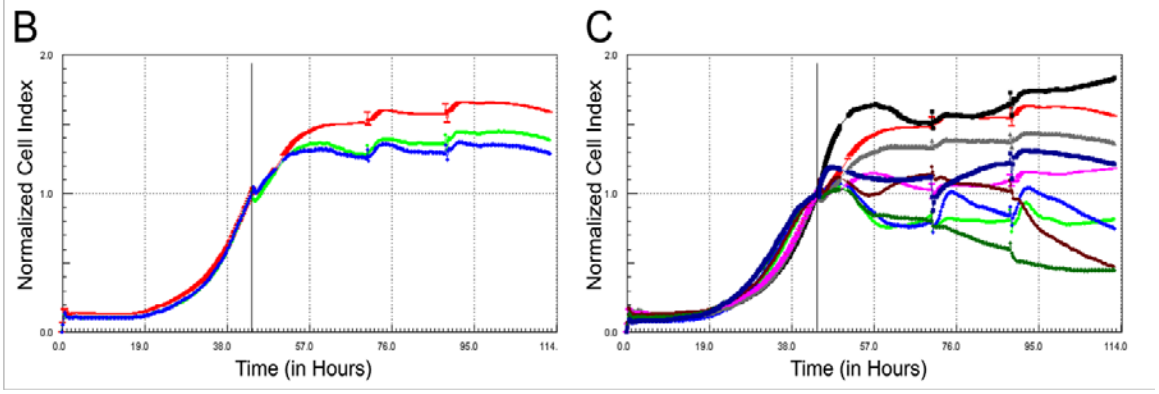


Figure 2-5. Real time impedance-based assay of affects of interfering with CK2 or PKC signaling in hESC.

(**A**) Characteristics of compounds and key. Normalized cell index of BG02 cultures treated with DMSO carrier control (red) and (**B**) two different CK2 inhibitors, or (**C**) PKC inhibitors, activators or a modulator. The traces are averages of 3 independent wells (CK2 and PKC effectors), or 6 wells (DMSO). Cell index plots were normalized at the time point prior to drug addition (48-hours post plating, vertical line). Transitory spikes in impedance are related to minor temperature fluctuations caused by daily feeding. Measurements were taken every 15 minutes, with an n incidental gap 6 hours after compound addition.

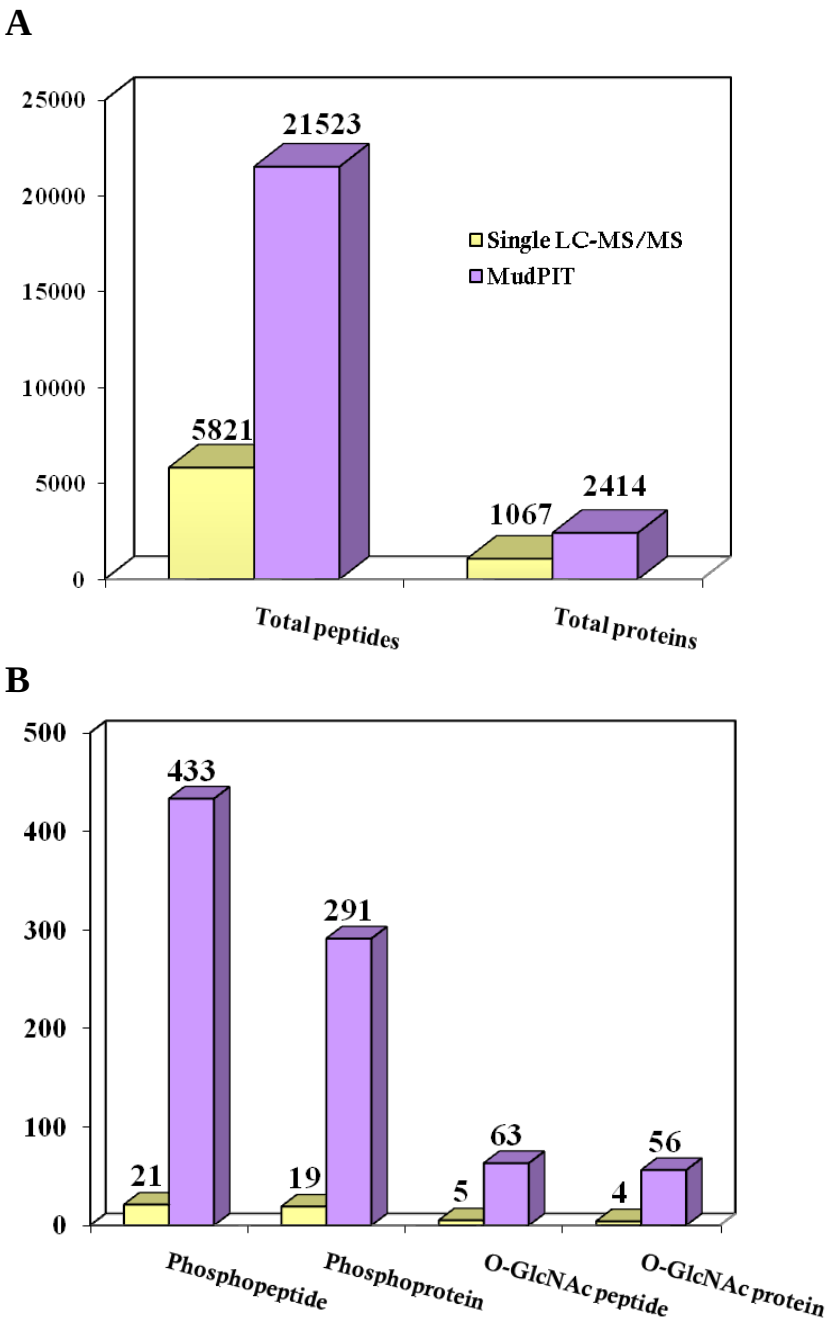


Figure 2-6. Comparison between single LC-MS/MS and MudPIT experiments.
(**A**) Comparison of the total numbers of identified proteins and peptides; (**B**) Comparison of the total numbers of the modified proteins and peptides. An overall 2.26 fold increase (2414/1067) in total number of the identified proteins was observed with a 3.7 fold increase (21523/5821) in the number of peptides in MudPIT compared to the single LC-MS/MS experiment. Both of phosphate and O-GlcNAc modified proteomes showed an increase for over an order of magnitude in the numbers of identified proteins and corresponding peptides in the MudPIT experiments.

Table 2-1A. Assignment of hES cell and secreted proteome (1% FDR).

	Membrane/ Organelle	Nucleus	Cytosol	Cytoskeleton	Combined hES Cell Proteome	Single-hit in Cell Proteome	Secreted Proteome
Protein	1844	2038	1677	852	3189	722	102
Peptide	18268	20902	17960	7439	24824	722	445
Spectral Count	68190	79573	84732	29101	261626	2162	4032

Table 2-1B. Assignment of PTMs within hES cell proteome (1% FDR).

	Protein	Peptide	Site	Novel Protein	Novel Site
Phosphorylation	288	432	492	31	200
O-GlcNAc Modification	55	63	68	37	68

Table 2-2. Novel phosphorylation sites and corresponding proteins identified in hES cell proteome.

Uniprot Accession	Protein Name	Site
A2A2P8	Adducin 1	S12[a]
O00232	26S proteasome non-ATPase regulatory subunit 12	T148[b], T151[b]
O43447	Peptidyl-prolyl cis-trans isomerase H	T159[a]
P05186	Alkaline phosphatase, tissue-nonspecific isozyme precursor	Y415[b]
P20073	Annexin A7	Y462[a]
P31040	Succinate dehydrogenase [ubiquinone] flavoprotein subunit,	T508[a], S509[a]
P33992	DNA replication licensing factor MCM5	Y565[a]
P33993	DNA replication licensing factor MCM7	S613[a], T614[a]
P35250	Replication factor C subunit 2	Y42[a]
P41250	Glycyl-tRNA synthetase	Y414[a], Y416[a], S423[a]
P48047	ATP synthase O subunit, mitochondrial precursor	S10[b], Y193[a]
P50991	T-complex protein 1 subunit delta	Y24[a]
P51648	Fatty aldehyde dehydrogenase	S14[a]
P53007	Tricarboxylate transport protein, mitochondrial precursor	Y276[a]
P61353	60S ribosomal protein L27	T58[a]
P62424	60S ribosomal protein L7a	Y181[a]
Q13423	NAD(P) transhydrogenase, mitochondrial precursor	Y37[a]
Q13492	Phosphatidylinositol-binding clathrin assembly protein	Y222[a]
Q16698	2,4-dienoyl-CoA reductase, mitochondrial precursor	T99[b]
Q2TAY7	Smu-1 suppressor of mec-8 and unc-52 protein homolog	Y132[a], S136[a], S137[a]
Q4VBZ6	EEF1D protein	S109[a]
Q59GV3	SWI/SNF-related matrix-associated actin-dependent regulator of	Y308[a]
Q5H9E3	THO complex 2	S1417[b]
Q5JR95	Ribosomal protein S8	Y168[a]
Q6DEN2	DPYSL3 protein	S636[a]
Q6IPL9	HMGA1 protein	T42[a]
Q6PGP7	Tetratricopeptide repeat protein 37	S733[a]
Q9BXA9	Sal-like protein 3	S39[b]
Q9H2R7	NPD011	S90[b], T96[a]
Q9HBD4	SMARCA4 isoform 2	S1449[b], S1602[b], S1607[b], S1659[b], S1663[b], S1676[b]
Q9UDR5	Alpha-aminoadipic semialdehyde synthase, mitochondrial precursor	Y203[a]
a	Novel (only one possible site of phosphorylation which is novel.)	
b	Novel (multiple possible sites of phosphorylation which are all novel.)	

Table 2-3. Implicated kinases predicted by Scansite with high-stringency filter.

Uniprot Accession	Name	Kinase			
Q9Y2W1	THRAP3	Cdc2, Cdk5, Akt			
Q58EY4	SMARCC1	Cdc2, Cdk5, GSK3			
Q5VVN3	SRRM1	Cdc2, Cdk5, Akt, Erk1, GSK3			
P11388	TOP2A	PKC_epsilon			
Q9HBD4	SMARCA4	CK2			
Q13242	SFRS9	Akt, Clk2			
Q8NC56	LEMD2	PKA			
Uniprot Accesstion	Protien Name	Motif found	Motif family	Site found	Site in dataset
Q9Y2W1	Thyroid hormone receptor-associated protein 3 (THRAP3)	Cdc2_Kin	Pro_ST_kin	S136	S134 S139
		Cdk5_Kin	Pro_ST_kin	S136	S134 S139
		Akt_Kin	Baso_ST_kin	S55	S55
Q58EY4	SWI/SNF-related matrix-associated actin-dependent	Cdc2_Kin	Pro_ST_kin	S310	S310
		Cdk5_Kin	Pro_ST_kin	S310	S310
		GSK3_Kin	Acid_ST_kin	T335	S330
Q5VVN3	Serine/arginine repetitive matrix protein 1 (SRRM1)	Cdc2_Kin	Pro_ST_kin	S393	S393
		Cdk5_Kin	Pro_ST_kin	S393	S393
		Cdc2_Kin	Pro_ST_kin	T406	T406
		Erk1_Kin	Pro_ST_kin	T406	T406
		GSK3_Kin	Acid_ST_kin	T406	T406
		Akt_Kin	Baso_ST_kin	T574	T572 T574
		Cdk5_Kin	Pro_ST_kin	T574	T572 T574
P11388	DNA topoisomerase 2-alpha (TOP2A)	PKC_epsilon	Baso_ST_kin	S1469	S1474
Q9HBD4	SMARCA4 isoform 2	Casn_Kin2	Acid_ST_kin	S1602	S1602
		Casn_Kin2	Acid_ST_kin	S1607	S1607
		Casn_Kin2	Acid_ST_kin	S1663	S1663
Q13242	Splicing factor, arginine/serine-rich 9 (SFRS9)	Akt_Kin	Baso_ST_kin	S199	S199
		Clk2_Kin	Baso_ST_kin	S199	S199
Q8NC56	LEM domain-containing protein 2 (LEMD2)	PKA_Kin	Baso_ST_kin	S134	S138 S139

Table 2-4. Novel O-GlcNAc modified proteins and corresponding sites.

Uniprot Accession	Protein Name	Site
O14818	Proteasome subunit alpha type 7	S93
O15320	Cutaneous T-cell lymphoma-associated antigen 5	S325
O76094	Signal recognition particle 72 kDa protein	T655
O94906	Pre-mRNA-processing factor 6	T828
O95777	U6 snRNA-associated Sm-like protein LSm8	T26
P00387	NADH-cytochrome b5 reductase 3	S30, T31
P04899	Guanine nucleotide-binding protein G(i), alpha-2 subunit	S6, S16
P0C0S5	Histone H2A.Z	S43
P12004	Proliferating cell nuclear antigen	S10
P25205	DNA replication licensing factor MCM3	T154, T617
P30520	Adenylosuccinate synthetase isozyme 2	T404, S413
P49406	39S ribosomal protein L19, mitochondrial precursor	S127
P49792	E3 SUMO-protein ligase RanBP2	T2192
P50991	T-complex protein 1 subunit delta	T217
P52701	DNA mismatch repair protein MSH6	S503
P61019	Ras-related protein Rab-2A	S121
P62495	Eukaryotic peptide chain release factor subunit 1	T388
P62910	60S ribosomal protein L32	S62
P78347	General transcription factor II-I	T785
Q08AD3	Nuclear pore complex protein Nup160	T306
Q14155	KIAA0142 splice variant 1	S642
Q52MB3	SAFB-like transcription modulator	S244
Q5C9Z4	Nucleolar MIF4G domain-containing protein 1	T167
Q5SWY0	Translocated promoter region	S1318
Q5VST9	Obscurin	T859
Q5VYK3	Proteasome-associated protein ECM29 homolog	T1375
Q6U8A4	Ubiquitin carboxyl-terminal hydrolase 7	T802
Q7KZ85	Transcription elongation factor SPT6	S125
Q7L7X3	Serine/threonine-protein kinase TAO1	T288
Q86YS8	BRD4-NUT fusion oncoprotein	T573
Q8NHH9	Atlastin-2	S308
Q8WXH0	Nesprin-2	S1090, S1102
Q92552	Mitochondrial 28S ribosomal protein S27	S349
Q96T37	Putative RNA-binding protein 15	T401
Q9BQG0	Myb-binding protein 1A	S1308
Q9NXF1	Testis-expressed sequence 10 protein	T31
Q9UHD8	Septin-9	T142

Table 2-5. Proteins identified as being related to pluripotency of hESCs.

Uniprot Accession	Gene Name	Protein Name	Total Spectra	Total Peptides	Unique Peptides	Subcellular Localization
Q01860	POU5F1	POU domain, class 5, transcription factor 1	88	24	7	Nucleus
P48431	SOX2	Transcription factor SOX-2	7	3	3	Nucleus
Q8NEB9	PIK3C3	Phosphatidylinositol 3-kinase catalytic subunit type 3	2	1	1	Cytoplasm
P40763	STAT3	Signal transducer and activator of transcription 3	7	3	2	Nucleus
P49841	GSK3B	Glycogen synthase kinase-3 beta	2	1	1	Cytoplasm
P28482	MAPK1	Mitogen-activated protein kinase 1	2	1	1	Cytoplasm
P07947	YES1	Proto-oncogene tyrosine-protein kinase Yes	6	2	2	Cytoplasm
Q7L190	DPPA4	Developmental pluripotency-associated protein 4	351	57	16	Nucleus
Q5T230	UTF1	Undifferentiated embryonic cell transcription factor 1	278	57	21	Nucleus
Q9UJU5	FOXD3	Forkhead box protein D3	1	1	1	Nucleus
Q9H9Z2	LIN28	Lin-28 homolog A	327	57	14	Extracellular

Table 2-6. Proteins that are modified by both phosphate and O-GlcNAc.

Uniprot Accession	Protein Name	Modification Site		Novel Protein (Y/N)	
		Phosphorylation	O-GlcNAc Modification	Phosphorylation	O-GlcNAc Modification
P02545	Lamin-A/C	S390 S392	T91	N	N
P13639	Elongation factor 2	T57	T657	N	N
P19338	Nucleolin	T59 S60 S67 T69 S563	T301 T438	N	N
P25205	DNA replication licensing factor MCM3	S672 T674 S711 T722	T154 T617	N	Y
P49327	Fatty acid synthase	T2204	T315	N	N
P49792	E3 SUMO-protein ligase RanBP2	T2128 S2280	T2192	N	Y
P50991	T-complex protein 1 subunit delta	Y24	T217	Y	Y
Q08AD3	Nuclear pore complex protein Nup160	T1080	T306	N	Y
Q5SWY0	Translocated promoter region	S1108	S1318	N	Y
Q8WZ42	Titin	S1465 Y2602 T2603 T2611 Y13346 T13690 Y17874 Y20867 S20868 Y23644 Y32869	S1571 T14674 S28157	N	N
Q96T37	Putative RNA-binding protein 15	S656 S670 S674	T401	N	Y
Q9BQG0	Myb-binding protein 1A	S1267	S1308	N	Y
Q9NR30	Nucleolar RNA helicase 2	S89 S121	S759	N	N
Q9UHD8	Septin-9	S85	T142	N	Y

Table 2-7. List of identified secreted proteins of hESCs.

Uniprot Accession	Gene Name	Protein Name	Functional Class	Subcellular localization	Total Spectra	Total Peptides	Unique Peptides	Category
P14174	MIF	Macrophage migration inhibitory factor	Cytokine	Extracellular	18	3	1	b
P43490	PBEF1	Nicotinamide phosphoribosyltransferase	Transferase	Extracellular	40	12	10	b
Q05707	COL14A1	Collagen alpha-1(XIV) chain precursor	Cell adhesion	Extracellular	33	4	1	b
Q9H9Z2	LIN28	Lin-28 homolog A	Translation regulator	Extracellular	327	57	14	b
P02649	APOE	Apolipoprotein E precursor	Transporter	Extracellular	104	40	12	b
P07942	LAMB1	Laminin subunit beta-1 precursor	Cell adhesion	Extracellular	16	10	10	b
P11047	LAMC1	Laminin subunit gamma-1 precursor	Cell adhesion	Extracellular	72	31	14	b
P21741	MDK	Midkine precursor	Growth factor	Extracellular	8	3	2	b
P22105	TNXB	Tenascin-X precursor	Cell adhesion	Extracellular	50	6	5	b
P23142	FBLN1	Fibulin-1 precursor	Cell adhesion	Extracellular	3	2	1	b
P39060	COL18A1	Collagen alpha-1(XVIII) chain precursor	Cell adhesion	Extracellular	22	8	4	b
P78504	JAG1	Jagged-1 precursor	Growth factor	Extracellular	2	2	1	b
Q5JTP4	LAMA5	Laminin, alpha 5	Cell adhesion	Extracellular	2	1	1	b
Q5VUM2	LAMA2	Laminin, alpha 2	Cell adhesion	Extracellular	8	3	3	b
Q6P4A8	FLJ22662	LAMA-like protein 1 precursor	Hydrolase	Extracellular	5	3	2	b
Q8NHP8	P76	LAMA-like protein 2 precursor	Hydrolase	Extracellular	15	1	1	b
Q92520	FAM3C	Protein FAM3C precursor	Cytokine	Extracellular	71	17	8	b
Q9NR23	GDF3	Growth/differentiation factor 3 precursor	Growth factor	Extracellular	3	2	2	b
P12109	COL6A1	Collagen alpha-1(VI) chain precursor	Cell adhesion	Extracellular	4	2	2	b
Q12904	SCYE1	Multisynthetase complex auxiliary component p43	Cytokine	Extracellular	43	13	6	b
Q14847	LASP1	LIM and SH3 domain protein 1	Transporter	Extracellular	81	19	10	b
Q8IWA5	SLC44A2	Choline transporter-like protein 2	Transporter	Extracellular	5	2	1	b
Q9HD20	ATP13A1	Probable cation-transporting ATPase 13A1	Transporter	Extracellular	68	25	11	b
Q13454	TUSC3	Tumor suppressor candidate 3		Extracellular	2	1	1	b
P02751	FN1	Fibronectin precursor	Cell adhesion	Extracellular	7	2	1	b
O00292	LFTY2	Left-right determination factor 2 precursor	Growth factor	Extracellular	40	14	10	a
O75610	LFTY1	Left-right determination factor 1 precursor	Growth factor	Extracellular	1	1	1	a
P02787	TRFE	Serotransferrin precursor	Transporter	Extracellular	143	12	8	a
P06733	ENOA	Alpha-enolase	Transcription regulator	Extracellular	99	37	11	a
A2A2D0	STMN1	Stathmin 1/oncoprotein 18		Cytoplasm	14	3	3	c
a	Extracellular proteins identified from secreted sample							
b	Extracellular proteins identified from intracellular sample							
c	Proteins identified from secreted sample that are not in top1000 hESC intracellular proteins							

Table 2-S1. Comparison of specific protein sequence coverages between single LC-MS/MS and MudPIT experiments.

	∑Unique Peptide in Protein	
	Fatty acid synthase	Nucleolin
Single LC-MS/MS	52	29
MudPIT	103	63
Increase	1.98	2.17

CHAPTER 3

CANCER BIOMARKERS DISCOVERED IN PANCREATIC DUCTAL FLUID USING A GLYCOPROTEOMIC APPROACH[1]

[1] Peng Zhao, Melody Perlman-Porterfield, Michael Tiemeyer, Michael Pierce, Haiyong Han, Lance Wells To be submitted to *Journal of Proteome Research*.

ABSTRACT

Pancreatic cancer is the 4th leading cause of cancer death in Europe and the U.S. with a 5-year survival rate of less than 5%. In 2010, there are 43,140 people estimated to be diagnosed with pancreatic cancer, and 36,800 will die from it. One of the major causes of high mortality rate and low survival rate is the absence of reliable biomarkers for early diagnosis.

Our research adopted a MS-based glycoproteomic approach to profile the differentially expressed proteins and carbohydrates in human pancreatic ductal fluid. The data presented here show novel methods for the analysis of glycans, proteins, and glycoproteins from a single fluid.

The preliminary results of our analysis demonstrates several protein and glycan biomarker candidates based on parallel experiments from pancreatic ductal fluid (PDF) from donor groups sorted on the following criteria: pancreatic cancer, intraductal papillary mucinous neoplasms (IPMN), pancreatitis, and normal pancreas.

INTRODUCTION

Pancreatic cancer (adenocarcinoma) is a highly aggressive human malignancy with a poor prognosis that shows profound resistance to extant treatments. Although accounting for only 3% of all cancers, this disease represents 6-7% of all cancer related deaths and is ranked as the fourth leading cause of cancer death. In the United States, approximately 43140 new cases will be diagnosed with pancreatic cancer in 2010, and 36800 deaths are estimated [1]

Pancreatic cancer remains one of the most difficult to treat due to late initial diagnosis and to intrinsic resistance to conventional treatments. Delays in diagnosis are often due to small cancers or the presence of non specific symptoms. About 50% of patients have distant disease at the time of diagnosis (locally advanced stage) and in 40% the tumor has spread (metastatic stage). Despite the advances in therapeutic strategies including surgical techniques as well as local and systemic adjuvant therapies, the overall survival in patients with pancreatic cancer remains dismal and has not improved substantially over the past 30 years. Median survival from diagnosis is typically around 3 to 6 months, and the 5-year survival rate is lower than 5%[1]. Risk factors have been identified [2], molecular pathogenesis has been elucidated, but advances in early detection and efficient treatments remain rather disappointing despite tremendous efforts.

Research over the last few years has identified certain genetic alterations associated with pancreatic cancer (Some gene expression profiling studies have suggested certain genes of potential diagnostic significance for pancreatic cancer), such as point mutations of K-ras occurring in 90% of cases, and inactivation of other tumor suppressor genes such as p53 and p16 [3-4]. Unfortunately to date these have not led to any clinical application,

therefore, there is no recommended screening test for this disease. The current clinical tumor markers for pancreatic cancer, CEA and CA 19-9, lack the appropriate sensitivity and specificity required for screening an asymptomatic population to aid early diagnosis. In this regard, the discovery of reliable biomarkers for the early diagnosis of pancreatic cancer is imperative. This would facilitate the development of strategies for therapeutic interventions and improved management of cancer patients. In contrast to genomic studies, proteomics focuses on the dynamic proteome in a state of constant flux due to various modifications and regulation. Thus, analysis of the proteome not only provides information relating to an abnormally-regulated gene, but also the extent of its expression at a specific time point. The aim of cancer proteomics commences with the comparison of proteomes from diseased and control (normal/healthy) samples in large scale to identify differentially expressed proteins or PTMs (up- or down-regulated) for further quantification and identification. Glycosylation, one of the most common protein post-translational modifications (PTM), is involved in many physiological functions and biological pathways. Altered protein glycosylation has been associated with a variety of human pathology, including cancer, inflammatory and degenerative diseases. In comparison with the native forms, glycosylated proteins exhibit higher specificity and sensitivity as disease biomarkers, and therefore are becoming important targets for disease diagnosis, prognosis and therapeutic response to drugs. The emerging technology of glycoproteomics, which focuses on glycoproteome analysis, is increasingly becoming an important tool for biomarker discovery. An in-depth, comprehensive identification of aberrant glycoproteins, and further, quantitative detection of specific glycosylation

abnormalities in a complex environment require a concerted approach drawing from a variety of techniques.

In this study, a MS-based glycoproteomic approach was adopted to profile the differentially expressed proteins and carbohydrates in human pancreatic ductal fluid (PDF). The data presented here show novel methods for the analysis of glycans, proteins, and glycoproteins from a single fluid. The preliminary results of our analysis demonstrated several protein and glycan biomarker candidates based on parallel experiments from pancreatic ductal fluid from donor groups sorted on the following criteria: pancreatic cancer, intraductal papillary mucinous neoplasms (IPMN), pancreatitis, and normal pancreas.

EXPERIMENTAL PROCEDURE

Pancreatic ductal fluid specimen

Pancreatic ductal fluid samples were collected and provided by The Translational Genomics Research Institute (Phoenix, AZ), along with matching serum and plasma samples. All the pancreatic ductal fluid samples were organized into four different diagnoses: pancreatic cancer, intraductal papillary mucinous neoplasms (IPMN), pancreatitis, and normal pancreas. For protein identification and quantification, 12 samples in total were analyzed (N=3 for respective diagnosis) in parallel experiments.

Proteomic experiments – Sample selection and protein concentration determination

Clear (no visible blood or bile contamination) pancreatic ductal fluid samples were thawed on ice and filtered by 0.2 μm spin columns (Nanosep). Protein concentration of all the samples was determined using the micro BCA protein assay kit (Pierce)

following manufacture instruction. Based on the calculated concentration, equal amount of proteins (~1mg) were used for the downstream experiments.

Proteomic experiments – Protein digestion

The fluid samples were reduced with 10 mM DTT for 1 h at 56 °C, alkylated (carboxyamidomethylated) with 55 mM iodoacetamide (Sigma) in dark for 45 min, and digested with trypsin (Promega) in 40 mM NH_4HCO_3 overnight at 37 °C. The digestion was quenched with 1% trifluoroacetic acid (TFA), and the resulting peptides were desalted with C18 spin columns (Vydac Silica C18, The Nest Group, Inc.) and dried down in Speed Vac.

Proteomic experiments – Protein fractionation (RP-LC)

Protein fractionation was performed by reverse phase liquid chromatography (RP-LC) using the Agilent 1100 series HPLC system (Agilent Technologies). Solvent A (0.1% TFA) and solvent B (0.085% TFA/80% acetonitrile) were used to develop a linear gradient consisting of 5 min at 95% solvent A, 60 min gradient at variable slope to 95% solvent B, 3 min at 95% solvent B, 1.5 min to 95% solvent A, and 4.5 min at 95% solvent A. Dried peptides were dissolved in solvent A and separated on a 2.1 x 250 mm silica-based C18 column (VYDAC) at a flow rate of 100 µl/min over the linear gradient. Eluted peptides were collected every 4 min, and subsequently combined into 5 fractions (F1, 15-32%; F2, 32-40%; F3, 40-45%; F4, 45-55%; and F5, 55-85%), desalted and dried as described above.

Reverse phase nanoLC-MS/MS analysis

Dried peptides from each fraction generated by RP-LC (12 x 5 in total) were resuspended in 0.5 µl of solvent B (0.1% formic acid/80% acetonenitrile) and 19.5 µl of

solvent A (0.1% formic acid) and loaded on a 75 μm x 105 mm C18 reverse phase column (packed in house, YMC GEL ODS-AQ120ÅS-5, Waters) by nitrogen bomb. Peptides were eluted directly into the nanospray source of an LTQ Orbitrap XL™ (Thermo Fisher Scientific) with a 140-min linear gradient consisting of 5-100% solvent B over 90-95 min at a flow rate of ~250 nl/min. In order to optimize the separation of peptides eluted into the mass spectrometer, gradients were expanded over a 70-min period in the appropriate region corresponding to each fraction collected from the previous offline RP-LC separation (F1, 4-30%; F2, 9-35%; F3, 15-42%; F4, 20-55%; and F5, 28-85%). The spray voltage was set to 2.0 kV and the temperature of the heated capillary was set to 200 °C. Full scan MS spectra were acquired from m/z 300 to 2000 with a resolution of 60000 at m/z 400 after accumulation of 1000000 ions (mass accuracy < 2 ppm). MS/MS events were triggered by the 6 most intense ions from the preview of full scan and a dynamic exclusion window was applied which prevents the same m/z value from being selected for 6 seconds after its acquisition. All 5 sub-fractions were analyzed in technical triplicates and data were acquired using Xcalibur® (ver. 2.0.7, Thermo Fisher Scientific). It has been shown that for protein identification in mixture by tandem mass spectrometry, certain number of repeated experiments is required to reach a reasonable completeness (all predicted proteins in the mixture being identified). From statistical results, triplicates in MS/MS experiment can discover approximately 95% of all predicted proteins in samples with relatively high complexity.[5]

Proteomic data analysis

The acquired MS/MS spectra were searched against UniProt human proteome database (58831 entries, updated at May 10, 2009) using SEQUEST (Bioworks 3.3,

Thermo Fisher Scientific) with the following settings: 50-ppm and 0.5-Da deviation were set for monoisotopic precursor and fragment masses, respectively; trypsin was specified as the enzyme; only fully tryptic peptide identifications were retained; a maximum of 3 missed cleavage sites, 3 differential amino acids per modification and 3 differential modifications per peptide were allowed; oxidized methionine (+15.9949 Da) and carbamidomethylated cysteine (+57.0215 Da) were set as differential modifications. All of the raw spectra were searched against both normal and reversed database under the same parameters, and all of the output files from SEQUEST search were filtered and grouped by different biological samples and replicates in ProteoIQ™. The cutoff value of peptides was set to an Xcorr of 0.5 and the minimum peptide length was set to 4 amino acids. For protein identification, false discovery rate was set to 1% at protein level and peptides matched to multiple proteins were excluded; for protein quantification, the 1% protein-level false discovery rate data was further filtered to achieve a 10% peptide-level false discovery rate, and only proteins that are identified by more than one peptide and in more than one biological sample were considered. The validated result was submitted to Gene Ontology (www.geneontology.org)[6] for protein subcellular localization and biological function annotation.

In order to compare the protein expression level across samples with different diagnosis, normalized spectral abundance factors (NSAF) were calculated for each protein that was commonly observed in all four diagnoses. In this approach, the spectral counts (SpC) of each protein in a given dataset were divided by its length (L) and normalized to the sum of SpC/L values in the given dataset[7-8]:

$$NSAF_x = \frac{(\frac{SpC}{L})_x}{\sum_{i=1}^{N}(\frac{SpC}{L})_x}$$

To further resolve shared peptides between protein isoforms, a strategy of using distribution factor in the calculation of NSAF was employed [9]:

$$dNSAF = \frac{uSpC + [(d)(sSpC)]}{uL + sL}$$

$$d = \frac{uSpC}{\sum uSpC}$$

According to the equations above, dNSAF is calculated where spectral counts from shared peptides are distributed among protein isoforms based on a distribution factor, d. Spectral counts from peptides uniquely mapping to a protein are denoted as "uSpC", while spectral counts from peptides shared between isoforms are labeled "sSpC". Protein amino acid lengths mapped to unique and shared peptides are denoted as "uL" and "sL", respectively.

Immunoblotting

Protein concentration of normal and cancer pancreatic juice samples was determined by micro BCA protein assay as describe above. Equal amount of proteins from normal and cancer samples (ranging from 2-8 μg for different antibodies) were separated by 4-20% Tris-HCl precast minigels (Bio-Rad), and semi-dry transferred to Immobilon-P transfer membrane (Millipore). The membranes were blocked with 5% BSA in TBST (TBS with 0.1% Tween 20), and probed with each antibody at 4 °C overnight as follows: 1:1000 dilution for REG1α (Abcam), REG1β (Abcam), and REG3α (Abnova) blots, and 1:2000 dilution for phospholipase A2 (Abcam) and pancreatic lipase-related protein 2 (Abnova) blots. After the addition of secondary antibodies conjugated to

horseradish peroxidase (HRP) at room temperature for 1 h, the final detection of HRP activity was performed using SuperSignal West Pico chemiluminescent substrates (Thermo Fisher Scientific). The films were exposed to CL-XPosure film (Thermo Fisher Scientific).

RESULTS

Protein Identification

After filtering and removing duplicates, the combined data set consists of 451 proteins identified by 2082 peptides corresponding to 59127 spectra, 60% (267/451) of which were identified by more than one peptide. Specifically, 136 proteins were identified by 775 peptides with 11635 spectra in normal samples; 163 proteins were identified by 769 peptides with 6645 spectra in pancreatitis samples; 149 proteins were identified by 831 peptides with 22641 spectra in IPMN samples; and 213 proteins were identified by 1094 peptides with 18206 spectra in pancreatic cancer samples (Table 3-1). All the identified proteins were submitted to Gene Ontology (www.geneontology.org) for subcelullar localization and biological function annotation. Based on the spectral counts assigned to each identified protein, the majority of the proteins are secreted proteases (81.26%) involved in proteolysis (51.87%) and metabolic process (29.40%) (Figure 3-1).

Protein Quantification

To evaluate the variation in protein expression across pancreatic juice samples with different diagnosis, the identified protein data set was further filtered to achieve a 10% peptide-level false discovery rate at 1% protein-level false discovery rate. After filtering, the resulting data set was examined manually to eliminate proteins that were only identified by one peptide or in only one patient. In the final quantified data set, a

total of 47 proteins were quantified with 590 peptides and 46172 spectra across three diagnoses and normal controls. Specifically, 22 proteins were quantified with 300 peptides and 8674 spectra in normal samples; 19 proteins were quantified with 215 and 3774 spectra in pancreatitis samples; 35 proteins were quantified with 414 peptides and 18632 spectra in IPMN samples; and 36 proteins were quantified with 422 peptides and 15092 spectra in pancreatic cancer samples.

By comparing the dNSAF values of proteins that were commonly observed in the samples from normal control and three various diagnoses, we were able to discover the differential expression of 22 proteins in our dataset (Figure 3-2, Table 3-2). As presented in Figure 3-2 and Table 3-2, in reference to normal controls, several proteins, such as REG1α, α-amylase, trypsin-1, chymotrypsinogen B, and glycoprotein GP2 isoform 1, showed significant elevation in IPMN and cancer samples. Several other proteins, such as pancreatic amylase, elastase 2A, 3B and 3A, carboxypeptidase A1, and pancreatic lipase-related protein 2, were downregulated in IPMN and cancer samples compared to normal controls. We also found several proteins that were uniquely expressed in IPMN and/or cancer samples on the quantifiable level (Table 3-3), such as REG1β, REG3α, CCDC132, phospholipase A2, elastase 2B, etc. As we re-examined the uniquely expressed proteins on the identifiable level, we discovered that even though some of those proteins were unique in IPMN and/or cancer samples on quantifiable level, they may be observed universally in the other biological samples on identifiable level (Table 3-4). For example, REG1β was only seen in two cancer patients on the quantifiable level, however, it was identified in patients with all three diagnoses and normal controls.

DISCUSSION

According to the quantitative proteomic results from 12 pancreatic ductal fluid samples ranging across four different diagnoses, there are several proteins that are distinctively upregulated or downregulated in pancreatic cancer and/or IPMN samples, and some proteins are uniquely expressed in the malignant samples. These proteins have the potential to serve as biomarkers for the diagnosis and/or prognosis of pancreatic cancer.

REG

Reg protein, also known as lithostathine or regenerating islet-derived protein, is a group of proteins that are translated from regenerating gene (REG) family which is mainly involved in the liver, pancreatic, gastric and intestinal cell proliferation or differentiation. Reg protein has four known isoforms which are O-glycosylated on Thr-5 with variable glycan chains [10]. Certain members of the Reg protein superfamily, such as REG4 and REG1α, have been related to gastric cancer, liver cancer, and pancreatic cancer [11-14].

In our dataset, REG1α was shown to be elevated in IPMN and cancer in comparison to normal and pancreatitis samples, REG1β was uniquely quantified in cancer samples, and REG3α was uniquely quantified in IPMN and cancer samples. To verify the result, immunoblotting was performed against these three proteins as an orthogonal method to mass spectrometry-based label-free quantification. In Figure 3-4A, multiple bands from 15 to 22 kDa showed elevated signal in cancer samples for REG1α, whereas there were only two bands at 15 and 19 kDa observed in the normal control. In Figure 3-4B and 3-4C, a similar pattern was observed in REG1β and REG3α blots with

the elevated signals and multiple bands present in cancer samples. The molecular weight heterogeneity of Reg proteins is a result of its various glycoforms[10], and the bands at 15 and 19 kDa correspond to the tryptic polypeptide, which is probably a generated by the endogenous proteolytic activity in pancreatic ductal fluid, and the native Reg protein. This observation suggested that the glycosylated Reg proteins were upregulated or even uniquely expressed in pancreatic cancer samples since the glycoforms were missing in the normal controls. Reg proteins especially its glycoforms can be considered as positive markers for pancreatic malignancy.

Phospholipase A2 (PLA2)

Phospholipase A2 (PLA2) is involved in lipid metabolism and therefore serves important roles in several cellular processes. Several studies have associated an elevated level of PLA2 with various form of human malignancy, such as breast, lung and prostate cancers [15-25].

In our dataset, PLA2 was quantified as a unique protein expressed in IPMN and cancer samples. On identifiable level, it was also observed in one pancreatitis patient and two normal controls. In Figure 3-4D, distinctive bands were observed at 32 kDa (immunogen molecular weight) and 16 kDa (human PLA2 sequence) in the cancer samples and were absent in the normal controls. Phospholipase A2 can be considered as a positive marker for pancreatic malignancy.

CCDC132

Coiled-coil domain-containing protein 132 was not only uniquely quantified but also identified in only two cancer patients in our dataset. There has not been any literature relating CCDC132 to human malignancy, however, the uniqueness of CCDC132 in

pancreatic ductal malignancies as shown in our experiments indicates its potential of being a diagnostic or prognostic marker.

Pancreatic lipase-related protein 2 (PLRP2)

Pancreatic lipase-related protein 2, which is the major colipase-dependent lipase in the pancreas [26], has been associated with indirect killing tumor cells by releasing unsaturated fatty acids which at elevated concentrations can cause apoptotic and necrotic death of tumor cells [27-34]. In another study to assess enzyme secretory capacity in the pancreatic diseases, pancreatic lipase was also shown to decrease in chronic pancreatitis and pancreatic cancer, suggesting its susceptibility to pancreatic dysfunction [35].

In our dataset, compared to normal controls, pancreatic lipase-related protein 1 was downregulated in IPMN and absent in cancer samples on quantifiable level. Pancreatic lipase-related protein 2 was absent in IPMN and significantly downregulated in cancer samples in reference to protein expression in normal samples. In Figure 3-4E, there were distinctively increased signals in normal samples compared to cancer samples observed at 37 kDa (immunogen molecular weight) and 52 kDa (human PNLIPRP2 sequence). The decreased expression of pancreatic lipase-related protein 2 can be considered as negative indicators for pancreatic disorders.

Elastase

A significant decreased elastase output in the duodenal aspirates during a pancreozymin secretin test was observed in chronic pancreatitis, pancreatic cancer, and liver cirrhosis patients when compared to normal controls [36]. Another study proved the expression of elastase 3A and its splicing forms in pancreatic duction carcinoma cells by reverse transcription-polymerase chain reaction (RT-PCR)[37].

In our dataset, elastase (2A, 3A, 3B) was downregulated in IPMN and cancer samples compared to normal samples. This finding is consistent with another quantitative proteomic study of protein expression profiles in pancreatic adenocarcinoma tissue using 2D gel electrophoresis techniques coupled with mass spectrometry[38]. The validation by immunoblotting was inconclusive because the underexpression of elastase 2A, 3A and 3B were obscured possibly by the overexpression of elastase 1 (data not shown). The increased level of elastase 1 in pancreatic cancer has been observed by several groups [39-44], and because of the sequence homogeneity shared in elastase protein family, the signal in our immunoblotting experiment could be revealing the overall expression of elastase family instead of one specific isoform.

Pancreatic amylase

An elevated serum amylase level, known as hyperamylasemia, has been associated with pancreatic and salivary diseases. The origin of serum amylase was elucidated in electrophoresis studies of normal serum [45], which showed that serum amylase consists of two main types: P-type amylase from the pancreas and S-type amylase from the salivary glands. It has been shown that the level of pancreatic amylase in rats with induced pancreatic cancer was significantly decreased and their histological findings also showed a decrease in zymogen content together with its total absence in some areas of malignant cells [46], suggesting that the original carcinogenic events were associated with a decrease in amylase initial activity.

In our dataset, salivary α-amylase was significantly upregulated while pancreatic α-amylase was downregulated in IPMN and cancer samples compared to normal controls and pancreatitis samples. It is possible that the increased secretion of salivary α-amylase

from salivary glands is compensating for the decreased pancreatic α-amylase secretion due to pancreatic dysfunction to maintain the serum amylase level or even cause hyperamylasemia. Pancreatic amylase can be considered as a negative marker for pancreatic malignancy.

CONCLUSION

In this study, a glycoproteomic approach was adopted to analyze pancreatic ductal fluid in search for potential biomarkers for pancreatic cancer. Based on the result from the proteomic experiments, several proteins, such as REG proteins, phospholipase A2, CCDC132, pancreatic lipase-related protein 2, elastase, and pancreatic amylase, either showed significant change in the expression level between cancer and normal samples or were uniquely observed in cancer samples, and can be considered as cancer biomarker candidates. It is also possible to target those potential protein biomarkers for the development of antibody-based drugs in treating the disease or monitoring its progress.

REFERENCES

1. Jemal, A.; Siegel, R.; Xu, J.; Ward, E., Cancer statistics, 2010. *CA Cancer J Clin* **2010,** 60, (5), 277-300.

2. Klapman, J.; Malafa, M. P., Early detection of pancreatic cancer: why, who, and how to screen. *Cancer Control* **2008,** 15, (4), 280-7.

3. Hruban, R. H.; van Mansfeld, A. D.; Offerhaus, G. J.; van Weering, D. H.; Allison, D. C.; Goodman, S. N.; Kensler, T. W.; Bose, K. K.; Cameron, J. L.; Bos, J. L., K-ras oncogene activation in adenocarcinoma of the human pancreas. A study of 82 carcinomas using a combination of mutant-enriched polymerase chain reaction analysis and allele-specific oligonucleotide hybridization. *Am J Pathol* **1993,** 143, (2), 545-54.

4. Goggins, M.; Offerhaus, G. J.; Hilgers, W.; Griffin, C. A.; Shekher, M.; Tang, D.; Sohn, T. A.; Yeo, C. J.; Kern, S. E.; Hruban, R. H., Pancreatic adenocarcinomas with DNA replication errors (RER+) are associated with wild-type K-ras and characteristic histopathology. Poor differentiation, a syncytial growth pattern, and pushing borders suggest RER+. *Am J Pathol* **1998,** 152, (6), 1501-7.

5. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004,** 76, (14), 4193-201.

6. Harris, M. A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G. M.; Blake, J. A.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J. T.; Hill, D. P.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J. M.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S.; Fisk, D. G.; Hirschman, J. E.; Hong, E. L.; Nash, R. S.; Sethuraman, A.; Theesfeld, C. L.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S. Y.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E. M.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.; Tonellato, P.; Jaiswal, P.; Seigfried, T.; White, R., The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **2004,** 32, (Database issue), D258-61.

7. Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P., Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* **2006,** 5, (9), 2339-47.

8. Zybailov, B. L.; Florens, L.; Washburn, M. P., Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. *Mol Biosyst* **2007,** 3, (5), 354-60.

9. Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L., Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal Chem* **2010,** 82, (6), 2272-81.

10. De Reggi, M.; Capon, C.; Gharib, B.; Wieruszeski, J. M.; Michel, R.; Fournet, B., The glycan moiety of human pancreatic lithostathine. Structure characterization and possible pathophysiological implications. *Eur J Biochem* **1995,** 230, (2), 503-10.

11. Sekikawa, A.; Fukui, H.; Fujii, S.; Ichikawa, K.; Tomita, S.; Imura, J.; Chiba, T.; Fujimori, T., REG Ialpha protein mediates an anti-apoptotic effect of STAT3 signaling in gastric cancer cells. *Carcinogenesis* **2008,** 29, (1), 76-83.

12. Yamagishi, H.; Fukui, H.; Sekikawa, A.; Kono, T.; Fujii, S.; Ichikawa, K.; Tomita, S.; Imura, J.; Hiraishi, H.; Chiba, T.; Fujimori, T., Expression profile of REG family proteins REG Ialpha and REG IV in advanced gastric cancer: comparison with mucin phenotype and prognostic markers. *Mod Pathol* **2009,** 22, (7), 906-13.

13. Usami, S.; Motoyama, S.; Koyota, S.; Wang, J.; Hayashi-Shibuya, K.; Maruyama, K.; Takahashi, N.; Saito, H.; Minamiya, Y.; Takasawa, S.; Ogawa, J.; Sugiyama, T., Regenerating gene I regulates interleukin-6 production in squamous esophageal cancer cells. *Biochem Biophys Res Commun* **2010,** 392, (1), 4-8.

14. Zhou, L.; Zhang, R.; Wang, L.; Shen, S.; Okamoto, H.; Sugawara, A.; Xia, L.; Wang, X.; Noguchi, N.; Yoshikawa, T.; Uruno, A.; Yao, W.; Yuan, Y., Up-regulation of REG Ialpha accelerates tumor progression in pancreatic cancer with diabetes. *Int J Cancer* **2010**.

15. Tribler, L.; Jensen, L. T.; Jorgensen, K.; Brunner, N.; Gelb, M. H.; Nielsen, H. J.; Jensen, S. S., Increased expression and activity of group IIA and X secretory phospholipase A2 in peritumoral versus central colon carcinoma tissue. *Anticancer Res* **2007,** 27, (5A), 3179-85.

16. Jiang, J.; Neubauer, B. L.; Graff, J. R.; Chedid, M.; Thomas, J. E.; Roehm, N. W.; Zhang, S.; Eckert, G. J.; Koch, M. O.; Eble, J. N.; Cheng, L., Expression of group IIA secretory phospholipase A2 is elevated in prostatic intraepithelial neoplasia and adenocarcinoma. *Am J Pathol* **2002,** 160, (2), 667-71.

17. Graff, J. R.; Konicek, B. W.; Deddens, J. A.; Chedid, M.; Hurst, B. M.; Colligan, B.; Neubauer, B. L.; Carter, H. W.; Carter, J. H., Expression of group IIa secretory phospholipase A2 increases with prostate tumor grade. *Clin Cancer Res* **2001,** 7, (12), 3857-61.

18. Yamashita, S.; Ogawa, M.; Sakamoto, K.; Abe, T.; Arakawa, H.; Yamashita, J., Elevation of serum group II phospholipase A2 levels in patients with advanced cancer. *Clin Chim Acta* **1994,** 228, (2), 91-9.

19. Yamashita, S.; Yamashita, J.; Sakamoto, K.; Inada, K.; Nakashima, Y.; Murata, K.; Saishoji, T.; Nomura, K.; Ogawa, M., Increased expression of membrane-associated phospholipase A2 shows malignant potential of human breast cancer cells. *Cancer* **1993,** 71, (10), 3058-64.

20. Yamashita, S.; Yamashita, J.; Ogawa, M., Overexpression of group II phospholipase A2 in human breast cancer tissues is closely associated with their malignant potency. *Br J Cancer* **1994,** 69, (6), 1166-70.

21. Yamashita, J.; Ogawa, M.; Sakai, K., Prognostic significance of three novel biologic factors in a clinical trial of adjuvant therapy for node-negative breast cancer. *Surgery* **1995,** 117, (6), 601-8.

22. Laye, J. P.; Gill, J. H., Phospholipase A2 expression in tumours: a target for therapeutic intervention? *Drug Discov Today* **2003,** 8, (15), 710-6.

23. Sved, P.; Scott, K. F.; McLeod, D.; King, N. J.; Singh, J.; Tsatralis, T.; Nikolov, B.; Boulas, J.; Nallan, L.; Gelb, M. H.; Sajinovic, M.; Graham, G. G.; Russell, P. J.; Dong, Q., Oncogenic action of secreted phospholipase A2 in prostate cancer. *Cancer Res* **2004,** 64, (19), 6934-40.

24. Dong, Q.; Patel, M.; Scott, K. F.; Graham, G. G.; Russell, P. J.; Sved, P., Oncogenic action of phospholipase A2 in prostate cancer. *Cancer Lett* **2006,** 240, (1), 9-16.

25. Denizot, Y.; Chianea, T.; Labrousse, F.; Truffinet, V.; Delage, M.; Mathonnet, M., Platelet-activating factor and human thyroid cancer. *Eur J Endocrinol* **2005,** 153, (1), 31-40.

26. D'Agostino, D.; Lowe, M. E., Pancreatic lipase-related protein 2 is the major colipase-dependent pancreatic lipase in suckling mice. *J Nutr* **2004,** 134, (1), 132-4.

27. Lima, T. M.; Kanunfre, C. C.; Pompeia, C.; Verlengia, R.; Curi, R., Ranking the toxicity of fatty acids on Jurkat and Raji cells by flow cytometric analysis. *Toxicol In Vitro* **2002,** 16, (6), 741-7.

28. Heimli, H.; Finstad, H. S.; Drevon, C. A., Necrosis and apoptosis in lymphoma cell lines exposed to eicosapentaenoic acid and antioxidants. *Lipids* **2001,** 36, (6), 613-21.

29. Finstad, H. S.; Myhrstad, M. C.; Heimli, H.; Lomo, J.; Blomhoff, H. K.; Kolset, S. O.; Drevon, C. A., Multiplication and death-type of leukemia cell lines exposed to very long-chain polyunsaturated fatty acids. *Leukemia* **1998,** 12, (6), 921-9.

30. Finstad, H. S.; Heimli, H.; Kolset, S. O.; Drevon, C. A., Proliferation and types of killing of leukemia cell lines by very long chain polyunsaturated fatty acids. *Lipids* **1999,** 34 Suppl, S107.

31. Cury-Boaventura, M. F.; Pompeia, C.; Curi, R., Comparative toxicity of oleic acid and linoleic acid on Raji cells. *Nutrition* **2005,** 21, (3), 395-405.

32. Cury-Boaventura, M. F.; Gorjao, R.; de Lima, T. M.; Newsholme, P.; Curi, R., Comparative toxicity of oleic and linoleic acid on human lymphocytes. *Life Sci* **2006,** 78, (13), 1448-56.

33. Cury-Boaventura, M. F.; Curi, R., Regulation of reactive oxygen species (ROS) production by C18 fatty acids in Jurkat and Raji cells. *Clin Sci (Lond)* **2005,** 108, (3), 245-53.

34. Cury-Boaventura, M. F.; Pompeia, C.; Curi, R., Comparative toxicity of oleic acid and linoleic acid on Jurkat cells. *Clin Nutr* **2004,** 23, (4), 721-32.

35. Hayakawa, T.; Kondo, T.; Shibata, T.; Kitagawa, M.; Ono, H.; Sakai, Y.; Kiriyama, S., Enzyme immunoassay for serum pancreatic lipase in the diagnosis of pancreatic diseases. *Gastroenterol Jpn* **1989,** 24, (5), 556-60.

36. Mizuno, R.; Hayakawa, T.; Noda, A., Elastase secretion in pancreatic disease. *Am J Gastroenterol* **1985,** 80, (2), 113-7.

37. Shimada, S.; Yamaguchi, K.; Takahashi, M.; Ogawa, M., Pancreatic elastase IIIA and its variants are expressed in pancreatic carcinoma cells. *Int J Mol Med* **2002,** 10, (5), 599-603.

38. Shen, J.; Person, M. D.; Zhu, J.; Abbruzzese, J. L.; Li, D., Protein expression profiles in pancreatic adenocarcinoma compared with normal pancreatic tissue and tissue affected by pancreatitis as detected by two-dimensional gel electrophoresis and mass spectrometry. *Cancer Res* **2004,** 64, (24), 9018-26.

39. Ventrucci, M.; Pezzilli, R.; Gullo, L.; Plate, L.; Sprovieri, G.; Barbara, L., Role of serum pancreatic enzyme assays in diagnosis of pancreatic disease. *Dig Dis Sci* **1989,** 34, (1), 39-45.

40. Iwase, K.; Miyata, M.; Yamaguchi, T.; Kawaguchi, T.; Tanaka, Y.; Matsuda, H., Pancreatic ductal cell carcinoma producing pancreatic elastase 1. *J Surg Oncol* **1993,** 54, (3), 199-202.

41. Achilles, K.; Bednarski, P. J., Quantification of elastase-like activity in 13 human cancer cell lines and in an immortalized human epithelial cell line by RP-HPLC. *Biol Chem* **2003,** 384, (5), 817-24.

42. Wu, D.; Qian, J. M.; Deng, R. X.; Jiang, W. J.; Chen, Y. J.; Liu, X. H.; Lu, X. H., Evaluating the role of serum elastase 1 in the diagnosis of pancreatic cancer. *Chin J Dig Dis* **2006,** 7, (2), 117-20.

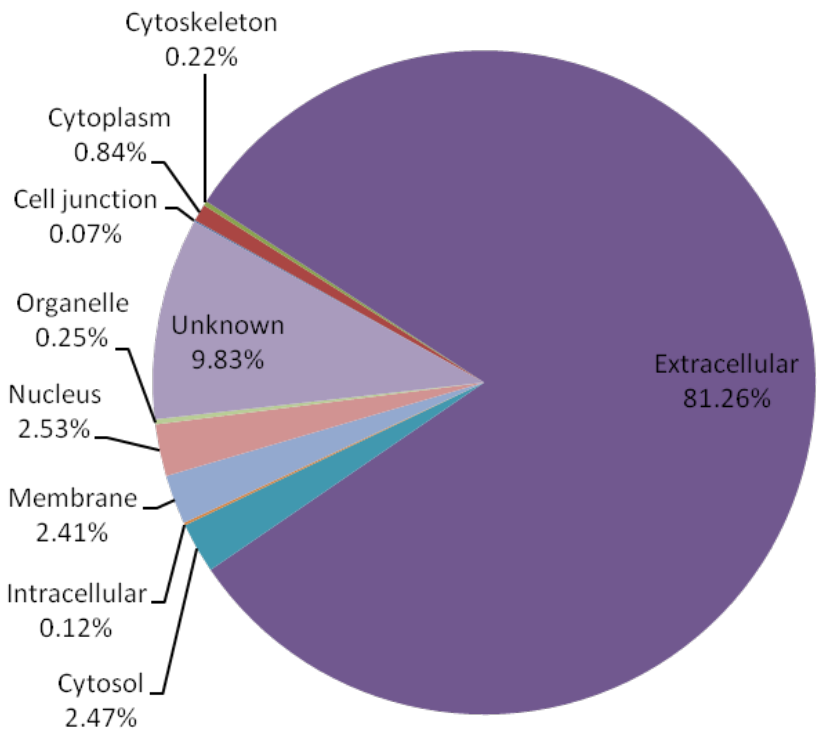
43. Wendorf, P.; Geyer, R.; Sziegoleit, A.; Linder, D., Localization and characterization of the glycosylation site of human pancreatic elastase 1. *FEBS Lett* **1989,** 249, (2), 275-8.

44. Wendorf, P.; Linder, D.; Sziegoleit, A.; Geyer, R., Carbohydrate structure of human pancreatic elastase 1. *Biochem J* **1991,** 278 (Pt 2), 505-14.

45. Pieper-Bigelow, C.; Strocchi, A.; Levitt, M. D., Where does serum amylase come from and where does it go? *Gastroenterol Clin North Am* **1990,** 19, (4), 793-810.

46. Hilmy, A. M.; Kandeel, K. M.; Selim, N. M., Pancreatic amylase as a tumour marker for pancreatic cancer. *Arch Geschwulstforsch* **1984,** 54, (6), 475-82.

A


Cytoskeleton
0.22%
Cytoplasm
0.84%
Cell junction
0.07%
Organelle
0.25%
Unknown
9.83%
Extracellular
81.26%
Nucleus
2.53%
Membrane
2.41%
Intracellular
0.12%
Cytosol
2.47%

B

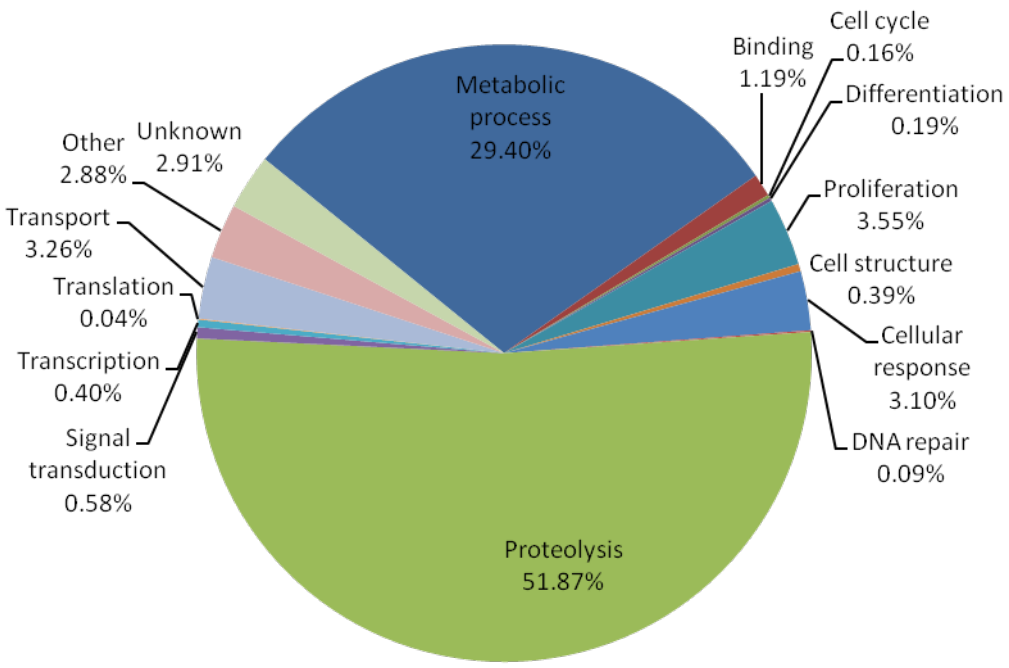


Figure 3-1. Biological annotation of the identified proteins in pancreatic ductal fluid samples.

Subcellular localization (**A**) and biological function (**B**) of identified proteins in 12 pancreatic ductal fluid samples. The distributions were calculated based on the spectral counts of identified proteins.

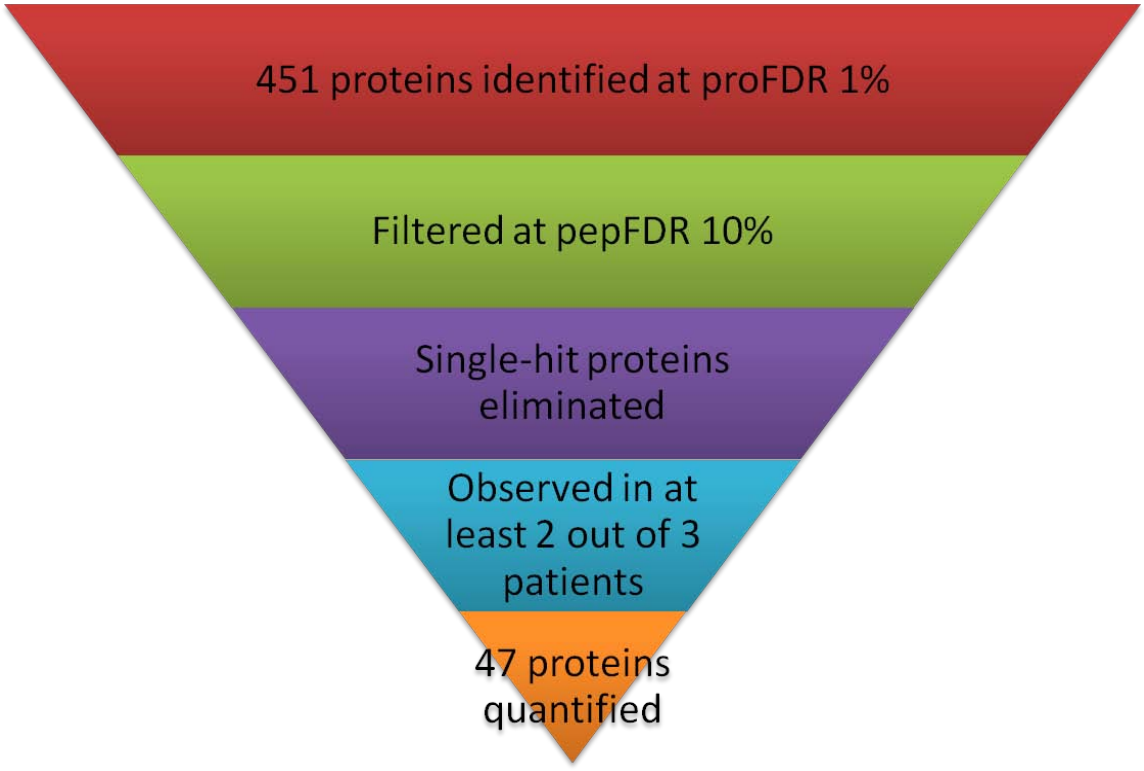
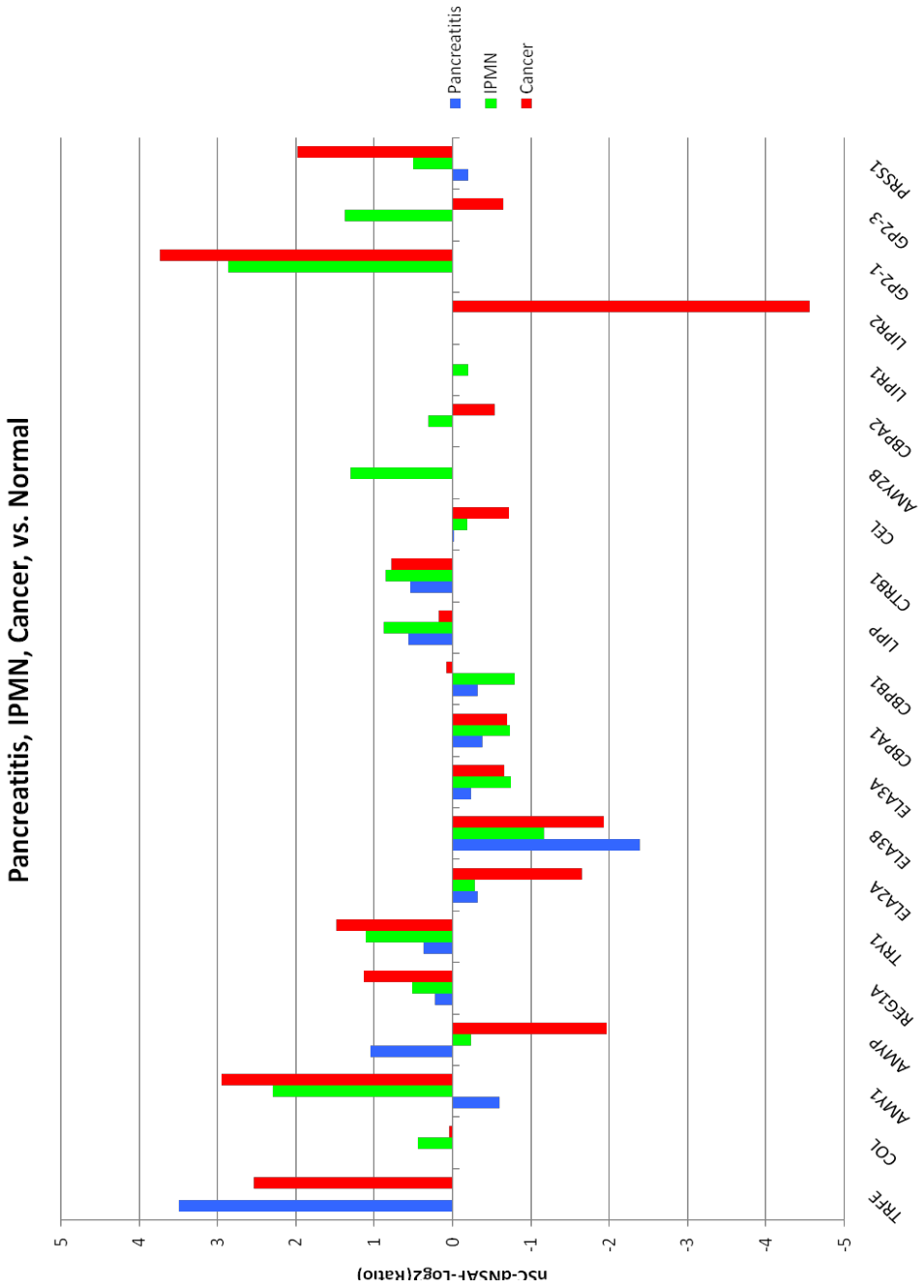


Figure 3-2. Data filter process flow chart.
451 proteins were identified by Sequest after filtering at 1% protein-level FDR. The dataset was then further filtered at 10% peptide-level 10% and 1-hit proteins were eliminated. In the resulting dataset, only proteins that were observed in at least 2 out of 3 patients were considered for quantitation, and finally 47 proteins were quantified.

A


Pancreatitis
IPMN
Cancer
Pancreatitis, IPMN, Cancer, vs. Normal
nSC-dNSAIt-Log2(Ratio)
5
4
3
2
1
0
-1
-2
-3
-4
-5
PRSS1
GP2-3
GP2-1
LIPR2
LIPR1
CBPA2
AMY2B
CEL
CTRB1
LIPP
CBPB1
CBPA1
ELA3A
ELA3B
ELA2A
TRY1
REG1A
AMYP
AMY1
COL
TRFE

B

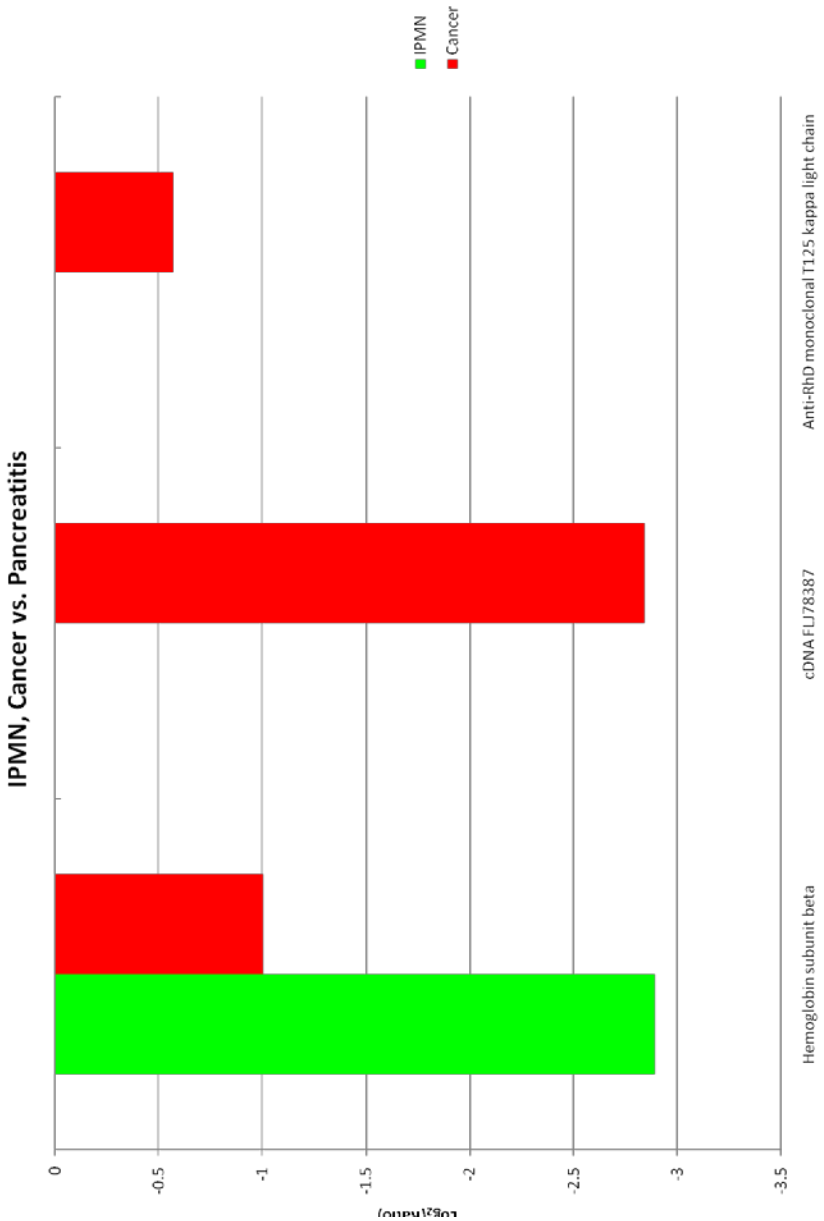


Figure 3-3. Variation of protein expression across pancreatic ductal fluid samples. (**A**) Protein expression variation in pancreatitis, IPMN, and cancer samples in reference to normal controls. The ratios were calculated based on the dNASF values of quantified proteins. (**B**) Protein expression variation in IPMN, and cancer samples in reference to pancreatitis samples. The ratios were calculated based on the dNASF values of quantified proteins.

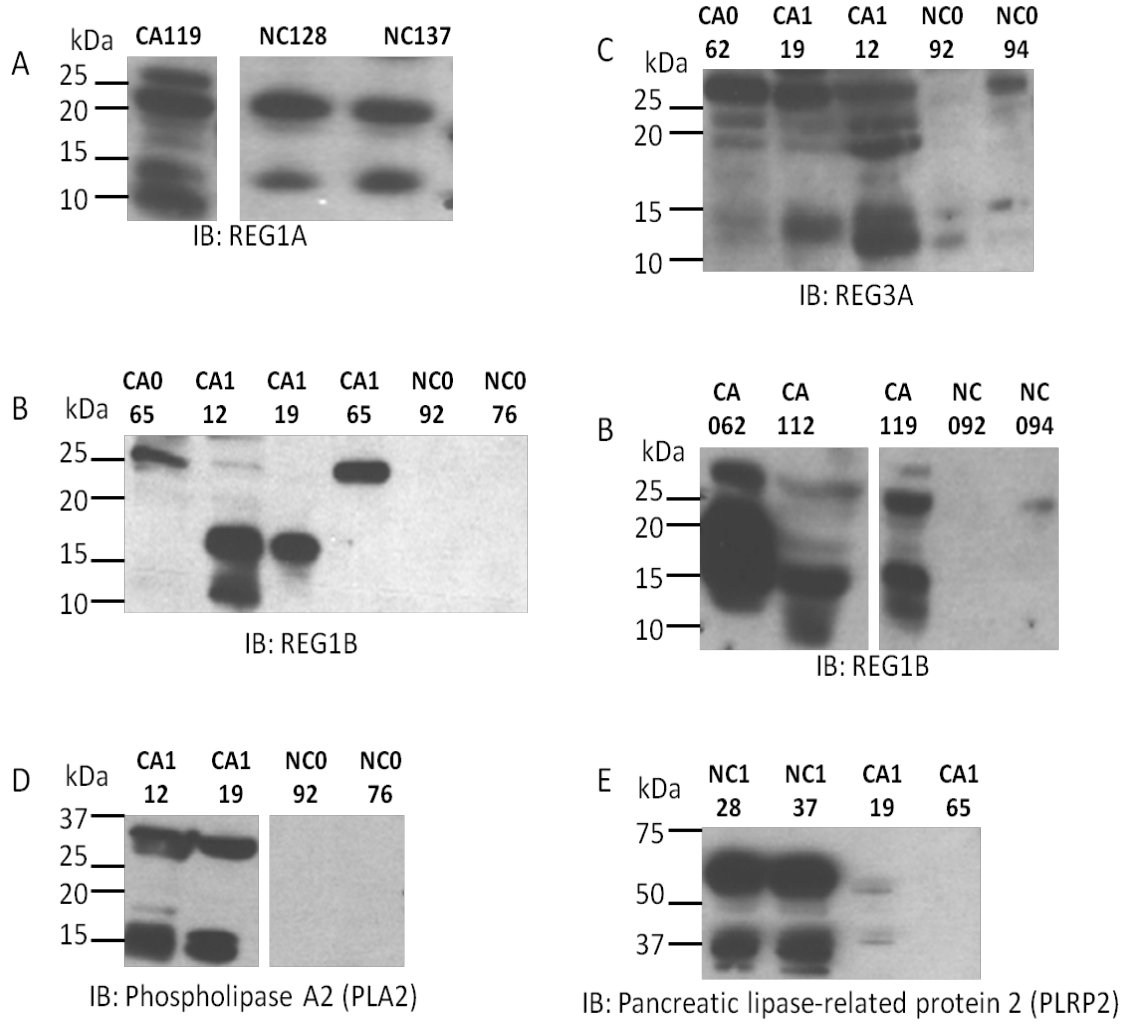


Figure 3-4. Validation of proteomic data by immunoblotting.
Several pancreatic ductal fluid samples with diagnosis of pancreatic cancer were compared to normal controls while probing with respective antibodies against (**A**) REG1α, (**B**) REG1β, (**C**) REG3α, (**D**) Phospholipase A2 (PLA2), and (**E**) Pancreatic lipase-related protein 2 (PLRP2). "CA" and "NC" indicate cancer samples and normal controls, respectively.

Table 3-1A. Summary of the identified proteins in pancreatic ductal fluid samples.

	Protein	Peptide	Spetral count	1-hit Assignment
Pancreatic Cancer	213	1094	18206	77
IPMN	149	831	22641	57
Pancreatitis	163	769	6645	61
Normal	136	775	11635	49
Total	451	2082	59127	184

Table 3-1B. Summary of the quantified proteins in pancreatic ductal fluid samples.

	Protein	Peptide	Spetral count	Unique Protein
Pancreatic Cancer	36	422	15092	7
IPMN	35	414	18632	8
Pancreatitis	19	215	3774	1
Normal	22	300	8674	0

Table 3-2A. Variation of protein expression levels in pancreatitis, IPMN, and cancer samples compared to normal controls.

Uniprot Accession	Protein Name	Gene Name	Ratio-PT/N	Log2ratio-PT/N	Ratio-IPMN/N	Log2ratio-IPMN/N	Ratio-CA/N	Log2ratio-CA/N
P02787	Serotransferrin	TF	11.24	3.49			5.81	2.54
P04118	Colipase	CLPS			1.35	0.44	1.03	0.04
P04745	Alpha-amylase 1	AMY1A	0.66	-0.60	4.92	2.30	7.73	2.95
P04746	Pancreatic alpha-amylase	AMY2A	2.07	1.05	0.85	-0.24	0.26	-1.96
P05451	Lithostathine-1-alpha	REG1A	1.17	0.23	1.43	0.52	2.19	1.13
P07477	Trypsin-1	PRSS1	1.29	0.37	2.16	1.11	2.79	1.48
P08217	Elastase-2A	ELA2A	0.80	-0.33	0.82	-0.29	0.32	-1.65
P08861	Elastase-3B	ELA3B	0.19	-2.39	0.44	-1.17	0.26	-1.93
P09093	Elastase-3A	ELA3A	0.85	-0.23	0.60	-0.75	0.63	-0.67
P15085	Carboxypeptidase A1	CPA1	0.77	-0.39	0.60	-0.73	0.62	-0.70
P15086	Carboxypeptidase B	CPB1	0.80	-0.32	0.58	-0.80	1.06	0.08
P16233	Pancreatic triacylglycerol lipase	PNLIP	1.48	0.57	1.84	0.88	1.13	0.18
P17538	Chymotrypsinogen B	CTRB1	1.45	0.54	1.81	0.86	1.72	0.78
P19835-1	Isoform Long of Bile salt-activated lipase	CEL	1.00	-0.01	0.88	-0.19	0.61	-0.72
P19961	Alpha-amylase 2B	AMY2B			2.47	1.30		
P48052	Carboxypeptidase A2	CPA2			1.24	0.31	0.69	-0.54
P54315-1	Isoform 1 of Pancreatic lipase-related protein 1	PNLIPRP1			0.87	-0.20		
P54317	Pancreatic lipase-related protein 2	PNLIPRP2					0.04	-4.56
P55259-1	Isoform 1 of Pancreatic secretory granule membrane major glycoprotein GP2	GP2			7.31	2.87	13.27	3.73
P55259-3	Isoform Alpha of Pancreatic secretory granule membrane major glycoprotein GP2	GP2			2.59	1.37	0.64	-0.65
Q3SY19	PRSS1 protein	PRSS1	0.87	-0.20	1.41	0.50	3.95	1.98
PT/N: Pancreatitis to Normal								
IPMN/N: IPMN to Normal								
CA/N: Cancer to Normal								

Table 3-2B. Variation of protein expression levels in IPMN and cancer compared to pancreatitis samples.

Uniprot Accession	Protein Name	Gene Name	Ratio-IPMN/PT	Log2ratio-IPMN/PT	Ratio-CA/PT	Log2ratio-CA/PT
P68871	Hemoglobin subunit beta	HBB	0.13	-2.89	0.50	-1.00
A8K008	cDNA FLJ78387				0.14	-2.84
Q5EFE6	Anti-RhD monoclonal T125 kappa light chain				0.67	-0.57
IPMN/PT: IPMN to Pancreatitis						
CA/PT: Cancer to Pancreatitis						

Table 3-3. Unique proteins observed in pancreatitis, IPMN and/or cancer on quantifiable and identifiable level.

Uniprot Accession	Protein Name	Gene Name	On Quantifiable Level		On Identifiable Level			
			In Cancer	In IPMN	In Cancer	In IPMN	In PT	In Normal
P01024	Complement C3	C3	2 patients	None	3 patients	None	1 patient	None
P01860	Ig gamma-3 chain C region	IGHG3	2 patients	None	2 patients	1 patient	None	None
P48304	Lithostathine-1-beta	REG1B	2 patients	None	2 patients	1 patient	1 patient (1hit)	1 patient (1hit)
Q96JG6-1	Isoform 1 of Coiled-coil domain-containing protein 132	CCDC132	2 patients	None	2 patients	1 patient	None	None
A0A5E4	Putative uncharacterized protein		2 patients	None	2 patients	1 patient	1 patient	1 patient (1hit)
Q569I7	Putative uncharacterized protein		2 patients	None	2 patients	1 patient	1 patient (1hit)	None
Q6ZP64	CDNA FLJ26451 fis, clone KDN03041		2 patients	None	2 patients	None	None	None
P04054	Phospholipase A2	PLA2G1B	2 patients	3 patients	3 patients	1 patient	2 patients	None
P08218	Elastase-2B	ELA2B	2 patients	3 patients	3 patients	1 patient	1 patient	None
P35030-1	Isoform A of Trypsin-3	PRSS3	2 patients	3 patients	3 patients	1 patient	1 patient	None
P69905	Hemoglobin subunit alpha	HBA1	2 patients	3 patients	3 patients	1 patient	None	None
Q06141	Regenerating islet-derived protein 3 alpha	REG3A	2 patients	2 patients	2 patients	None	None	None
Q99895	Chymotrypsin-C	CTRC	2 patients	3 patients	3 patients	1 patient	1 patient	None
PT: Pancreatitis								

CHAPTER 4

COMBING HIGH ENERGY COLLISION-INDUCED DISSOCIATION AND ELECTRON TRANSFER DISSOCIATION FOR PROTEIN O-GLCNAC MODIFICATION SITE ASSIGNMENT [1]

[1] Peng Zhao, Rosa Viner, Chin Fen Teo, David Horn, Lance Wells
Submitted to *Journal of Proteome Research*.

ABSTRACT

Mass spectrometry-based studies of proteins that are post-translationally modified by O-linked β-N-acetylglucosamine (O-GlcNAc) are challenged in effectively identifying the sites of modification while simultaneously sequencing the peptides. Here we tested the hypothesis that a combination of high-energy collision dissociation (HCD) and electron transfer dissociation (ETD) could specifically target the O-GlcNAc modified peptides and elucidate the amino acid sequence while preserving the attached GlcNAc residue for accurate site assignment. By taking advantage of the recently characterized O-GlcNAc-specific IgG monoclonal antibodies and the combination of HCD and ETD fragmentation techniques, O-GlcNAc modified proteins were enriched from HEK293T cells and subsequently characterized using the LTQ Orbitrap Velos™ ETD (Thermo Fisher Scientific) mass spectrometer. In our dataset, 58 sites of O-GlcNAc modification are reported confirming that the HCD/ETD combined approach is amenable to the detection and site assignment of O-GlcNAc modified peptides. Realizing HCD triggered ETD fragmentation on a linear ion trap/Orbitrap platform for more in-depth analysis and application of this technique to other post-translationally modified proteins are currently underway.

INTRODUCTION

Glycosylation through serine and threonine by a single O-linked β-N-acetylglucosamine (O-GlcNAc) moiety is a widespread post-translational modification seen in cytosolic and nuclear proteins. O-GlcNAc modification is a nutrient/stress-sensitive modification that regulates proteins involved in a wide array of biological processes, including transcription, signal transduction, and metabolism[1-2]. Cycling of O-GlcNAc is regulated by the concerted actions of O-GlcNAc transferase (OGT) and O-GlcNAcase (OGA)[3], and the fluctuation of O-GlcNAc levels has been associated with the etiology of type II diabetes, cardiovascular disease, and neurodegenerative disease[4-7]. Elucidating the molecular structure of O-GlcNAc modified proteins not only is crucial in revealing their site-specific functional roles but also is necessary in facilitating further discovery of the involvement of O-GlcNAc in major biological networks.

In order to compensate for the substoichiometric occupancy of O-GlcNAc modification[3, 8-9], numerous techniques have been developed for the detection and enrichment of O-GlcNAc modified proteins, such as immunoblotting[10-12], lectin affinity chromatography[13-14], and chemoenzymatic approach[15-16]. Facilitated by the advances in analytical technology, the identification of O-GlcNAc modified proteins following specific enrichment techniques are mostly accomplished by tandem mass spectrometry. However, since some of the enrichment techniques are performed on protein level, O-GlcNAc modified peptides often remain underrepresented in a proteolyzed mixture during a "bottom-up" proteomic experiment. Furthermore, due to the susceptibility of β-O-glycosidic bond to gas-phase collisional fragmentation[17-19], the GlcNAc residue is readily cleavable under collision-induced dissociation (CID) generating dominant neutral

loss ions, which suppresses the production of peptide backbone fragments and renders the site of modification unknown. Therefore, a reliable characterization of O-GlcNAc modified proteins, especially its molecular details including the complete sequence of proteins and the sites of modification, cannot be easily achieved in typical CID-oriented MS experiments. To date, proteomic analyses have identified more than 700 O-GlcNAc modified proteins in diverse functional classes[5], however, only a small percentage (<12%) of those proteins were assigned with modification sites. Recent advances in mass spectrometry, such as the introduction of high-energy collision dissociation (HCD) and the frequent use electron transfer dissociation (ETD) in addition to the most commonly used CID in proteomic studies, has provided us with the capability to perform unambiguous characterization of peptides with various modifications. When applied to post-translationally modified peptides, HCD tend to generate characteristic immonium or oxonium ions at low m/z region, such as phosphotyrosine immonium ion[20] and HexNAc oxonium ion[18-19, 21], which can serve as diagnostic tools for certain types of modification, whereas ETD produces sufficient c- and z-ions for confident peptide sequencing while preserving the modification residue attached for accurate site assignment.

In our study, by taking advantage of the recently characterized O-GlcNAc-specific IgG monoclonal antibodies[10] and the combination of HCD and ETD fragmentation techniques, a total of 58 O-GlcNAc modification sites survived the high-stringency filter and were successfully assigned on 16 proteins enriched from HEK293T cells, confirming that the HCD/ETD combined approach is amenable to the detection and site localization of O-GlcNAc modified peptides. Applicability of the HCD/ETD

approach to other type of glycosylated peptides, such as O-Mannose and O-GalNAc modified peptides, were also investigated in our study.

EXPERIMENTAL PROCEDURE

Monoclonal antibodies

The three monoclonal antibodies used in our current study, 18B10.C7(3), 9D1.E4(10) and 1F5.D6(14), were generated and characterized as described in the previous study[10].

Standard and synthetic peptides

In our experiments, three O-GlcNAc modified standard peptides were used, which are CREB [TAPTs(GlcNAc)TIAPG], CKII [PGGSTPVs(GlcNAc)SANMM], and BPP [PSVPVs(GlcNAc)GSAPGR]; and three O-Mannose and O-GalNAc modified synthetic peptides were used, which are Ac-IRt(Man)t(Man)t(GalNAc)SGVPR, Ac-PTTt(GalNAc)PLK, and Ac-RIRTT t(Man)SGVPR.

Protein digestion

1 mg of bovine serum albumin (Fisher) was incubated with 10 mM DTT and 40 mM NH_4CO_3 at 56 °C for 1 hr. After cooling to room temperature, the suspension was incubated for 45 min in dark with 55 mM iodoacetamide dissolved in 40 mM NH_4CO_3. After denaturing and alkylation, the sample was digested overnight at 37 °C using sequencing grade modified trypsin (Promega). The reaction was quenched by 0.1% trifluoroacetic acid, and the resulting peptides were divided into 4 aliquots (250 µg each), desalted using Vydac C18 Silica spin columns (Nest Group) and dried in SpeedVac[22].

Preparation of multiple-antibody-enriched HEK293T cell extract

The O-GlcNAc proteome enrichment from HEK293T cell pellets were prepared as previously described by Teo et al.[10]. Briefly, HEK293T cells were obtained from ATCC and maintained in Dulbecco's modified Eagle's medium (4.5 g/L glucose; Cellgro/Mediatech, Inc.) supplemented with 10% fetal bovine serum (GIBCO/Invitrogen) in 37 °C incubator humidified with 5% CO_2. Cells were harvested 48 h post-transfection, treated with 50 μM of PUGNAc for 24 h, pelleted and stored at -80 °C until used. The nucleocytosolic fraction of PUGNAc-treated HEK293T cell lysates were prepared by using hypotonic buffers and centrifugation. After preclearing with a mixture of normal mouse IgG AC and protein A/G PLUS agarose, the PUGNAc-treated HEK293T nucleocytosolic fractions were immunoprecipitated by conjugated MAbs 18B10.C7(3), 9D1.E4(10) and 1F5.D6(14), subsequently eluted with glycine (0.1 M, pH 2.5), and immediately neutralized with Tris-HCl (1 M, pH 8.0). The samples were then reduced and alkylated as described above and subjected to LysC (Roche) digestion at 37 °C for 24 h. After digestion, the samples were desalted and dried as described above.

LC-MS/MS analysis of peptide mixtures

Two peptide mixtures were analyzed in the experiment. The first mixture was produced by mixing 2 nmol of each O-GlcNAc modified standard peptides (sequence as described above) with equal molar of bovine serum albumin digest. The second mixture was produced by mixing equal molar of three synthetic O-Mannose and O-GalNAc modified peptides (sequence as described above). Each peptide mixture was resuspended in 1 μl of solvent B (0.1% formic acid/80% acetonenitrile) and 39 μl of solvent A (0.1% formic acid) and loaded on a 75 μm i.d. x 105 mm C18 reverse phase column (packed in

house, YMC GEL ODS-AQ120ÅS-5, Waters) by nitrogen bomb. Peptides were eluted directly into the nanospray source of an LTQ Orbitrap XL™ (Thermo Fisher Scientific) with a 160-min linear gradient consisting of 5-100% solvent B over 100 min at a flow rate of ~250 nl/min. The spray voltage was set to 2.0 kV and the temperature of the heated capillary was set to 200 °C. Full MS scans were acquired from m/z 150 to 2000 at a resolution of 60000 (FWHM at m/z 400), with a maximum ion injection time of 100 ms, and an automatic gain control (AGC) setting of 700000 ions. AGC was set to 30000 ions for MS/MS analysis (CID and ETD modes) in the ion trap and to 500000 ions for the MS/MS analysis (HCD mode) in the Orbitrap. The HCD normalized collision energy was set to 35%, and fragment ions were detected in the Orbitrap at a resolution of 7500 (FWHM at m/z 400) using 1 microscan, with a maximum injection time of 100 ms. For ion trap CID MS/MS, isolation of 2 amu, 1 microscan with a maximum injection time of 100 ms were used; for ion trap ETD MS/MS, isolation of 2 amu, 1 microscan with a maximum injection time of 300 ms were used. ETD fragmentations were performed based on charge state with the anion AGC target set at 300000. A dynamic exclusion window was applied which prevents the same m/z value from being selected for 6 seconds after its acquisition. Data acquisition was conducted in the fashion of an Orbitrap MS followed by top 4 data-dependent Orbitrap HCD MS/MS, ion trap ETD MS/MS, and ion trap CID MS/MS triple-play using Xcalibur® (ver. 2.0.7, Thermo Fisher Scientific).

LC-MS/MS analysis of enriched HEK293T cell extract

Three LC-MS/MS experiments were performed on an LTQ Orbitrap Velos™ ETD mass spectrometer (Thermo Scientific) with nano-ESI source that was coupled to a Surveyor™ MS Pump with a flow splitter. Peptides were separated on a 75 μm i.d. x 200

mm spraytip Magic C18 column (Michrom Bioresources) with a gradient elution from 5-20% over 120 min and from 20-40% over 70 min at a flow rate of ~300 nL/min using acetonitrile in 0.1% formic acid. The LTQ Orbitrap Velos™ mass spectrometer was operated at S-lenses setting of 50 %, the heated capillary temperature of 200 °C, resolution of 60000 (FWHM at m/z 400) in full MS, with a maximum ion injection time of 300 ms, and AGC setting of 1000000 ions. AGC was set to 10000 ions for MS/MS analysis in the ion trap and to 50000 ions for the MS/MS analysis in the Orbitrap. HCD normalized collision energy was set to 35% and fragment ions were detected in the Orbitrap at a resolution of 7500 (FWHM at m/z 400) using 1 microscan, with a maximum injection time of 200 ms. ETD reactions were performed based on charge state with the anion AGC target set at 200000. Three different LC-MS/MS acquisition methods were performed: (1) Orbitrap MS followed by top 10 data-dependent ion trap ETD MS/MS; (2) Orbitrap MS followed by top 5 data-dependent Orbitrap HCD MS/MS and ion trap ETD MS/MS double-play; (3) Orbitrap MS followed by top 5 data-dependent Orbitrap HCD MS/MS. Data were acquired using Xcalibur® (ver. 2.0.7, Thermo Fisher Scientific).

Data analysis

For the peptide mixtures, the raw spectra were interpreted manually. For the enriched HEK293T sample, the database searches were performed allowing for the same parameters as described above except for the specified enzyme as LysC and the differential modification settings. Each raw spectra file was searched twice in the following fashion (Figure 4-1): the first search, allowing oxidized methionine (+15.995 Da) and carbamidomethylated cysteine (+57.021 Da) as differential modifications; the

second search, allowing oxidized methionine (+15.995 Da), carbamidomethylated cysteine (+57.021 Da), and HexNAc modified serine/threonine (203.079 Da) as differential modifications. The output from the first search was filtered at 1% false discovery rate on the peptide level, and single-hit protein assignments were eliminated to generate the list of identified proteins; the output from the second search was filtered at 10% false discovery rate on the peptide level, and only the ones that were identified in both searches were retained as HexNAc-modified proteins. All the HexNAc modified peptides were manually validated.

A multiple-engine search strategy was applied to the same raw spectra files obtained from the HEK293T cell sample to increase the sensitivity of data analysis. In this strategy (Figure 4-S1), the database searches were performed using the same database described earlier with three search engines: SEQUEST® (Thermo Proteome Discoverer ver. 1.1.0.263, Thermo Fisher Scientific), Mascot® (www.matrixscience.com), and ProteinProspector (v5.6.1, prospector.ucsf.edu). The SEQUEST search comprised of two consecutive sections: the first search allowed for N-terminal and cysteine carbamidomethylation (+57.021 Da), methionine oxidation (+15.995 Da) and dethiomethylation (-48.003 Da); the second search used the protein dataset generated by the first search as the database and allowed for serine/threonine HexNAc modification (+203.079 Da) and serine/threonine/tyrosine phosphorylation (+79.966 Da) in addition to the 3 differential modifications set-up in the first search. Both sections were searched with: 20-ppm tolerance for monoisotopic precursor mass, and 1.2-Da or 0.01-Da tolerance for monoisotopic fragment masses of ion trap MS/MS or Orbitrap MS/MS spectra, respectively; LysC (fully digested) specified as the enzyme; a

maximum of 3 missed cleavage sites, 3 differential amino acids per modification and 3 differential modifications per peptide allowed. The list of identified proteins was generated only from the second SEQUEST database search and was filtered to achieve a 1% peptide-level false discovery rate. The list of HexNAc modified peptides was the combination of multiple-engine search results and each modification site was manually verified.

RESULTS AND DISCUSSION

The MS scheme employed in this study for sequencing O-GlcNAc modified peptides and assigning the modification sites is a combination of HCD and ETD fragmentation methods. When analyzing O-GlcNAc modified peptides, the unique advantage of HCD fragmentation is the generation of distinct HexNAc oxonium ions (m/z 204.09)[18-19, 21] and a series of HexNAc fragments (m/z 186, m/z 168, m/z 144, m/z 138 and m/z 126) which are significant to the diagnosis of O-GlcNAc modified peptides yet not necessarily observed in ion trap tandem mass spectra because of the dependence of ion trap scan range on precursor m/z values. Moreover, the product ions formed during HCD are detected in the Orbitrap analyzer, therefore not only do they overcome the limitation of 1/3 cut-off, they also exhibit higher mass accuracy and lower chemical noise. However, HCD spectra alone are not as informative as those acquired in the linear ion trap pertaining peptide sequence-related ion production, which has been attributed to the increased collision energy leading to ion scattering and a lack of peptide backbone fragments[23].

In order to provide more informative spectra for peptide sequencing as well as the site assignment of O-GlcNAc modification, ETD was employed in combination to HCD.

By transferring an electron from a radical anion to a protonated peptide, ETD has been proven advantageous for analyzing relatively large, non-tryptic peptides compared to CID. In the particular case of post-translationally modified peptides, ETD has the capacity to preserve labile modifications attached to peptide backbones, such as phosphorylation and glycosylation, allow for the detection of multiple modifications within the context of one another, as well as produce almost complete series of peptide backbone fragments for peptide sequencing in the meantime[24]. Especially when facilitated by supplementary collisional activation converting the nondissociative electron transfer products into c- and z- type fragment ions, the effect of precursor ion charge states have on dissociation efficiency, which is one of the limitations imposed by the intrinsic charge-reducing process in electron-based fragmentation methods, has been remarkably reduced[25].

By utilizing the diagnostic ions generated by HCD and the extensive peptide sequence information provided by ETD, we analyzed glycosylated standard peptides to investigate the applicability of HCD/ETD method in identification and site localization of O-GlcNAc modified peptides. Furthermore, we applied the same MS scheme to a complex biological sample to prove the effectiveness and robustness of this method.

Characterization of O-GlcNAc modified standard peptides

O-GlcNAc modified standard peptides in the mixture were analyzed by an alternating CID/ETD/HCD fragmentation method. As indicated in Figure 4-2, the most dominant peaks formed during CID fragmentation of glycopeptides CKII and BPP resulted from the loss of HexNAc residue from respective precursor ions (Figure 4-2A and 4-2E). As a result of the intensive HexNAc-loss ions, b- and y-type ions generated by

backbone fragmentation that are required for peptide sequencing and site assignment are severely suppressed, making it difficult to accurately and confidently characterize O-GlcNAc peptides.In Figure 4-2B and 4-2F, almost a complete series of c- or z-type ions were observed in the ETD spectrum allowing for the elucidation of both respective peptide sequence and modifications sites.During the HCD fragmentation of CKII and BPP glycopeptides (Figure 4-2C and 4-2G), the high m/z fragment ions were either missing or at very low intensity. However, the HexNAc oxonium ion (m/z 204) and its fragments (m/z 186, m/z 168, m/z 144, m/z 138, and m/z 126) were produced at pronouncedly high intensity (Figure 4-2D and 4-2H, Table 4-S1). This distinctive fragment pattern of HexNAc residue allows for unambiguous identification of O-GlcNAc modified peptides which are generally underrepresented in protein mixtures extracted from biological samples[3, 8-9].

When O-GlcNAc modified peptides are subjected to CID, it is probable, yet not necessary, that the GlcNAc residue is cleaved from the peptide backbone and forms an oxonium ion at m/z 204.09. When undergoing HCD, the GlcNAc residue will always be cleaved and generate the oxonium ion at m/z 204.09. This characteristic ion product can be used as a diagnostic ion to identify the HexNAc modified peptides. Moreover, not only will the HexNAc oxonium ion be present in the HCD spectrum, a series of its fragment ions will also be produced that are indicative of HexNAc residue, such as m/z 186, m/z 168, m/z 144, m/z 138, and m/z 126[26]. The presence of HexNAc oxonium ion and the series of its fragment ions can be used to target HexNAc modified peptides during an MS^n experiment and selectively trigger an equal order CID or ETD fragmentation of the same precursor ion for more detailed peptide characterization.

Compared to the dominant neutral loss ions created upon the loss of HexNAc residues during CID, which suppress the formation of peptide sequencing-related ions, ETD has the capability to preserve labile modifications and render more complete ion series for peptide sequencing and site assignment. Using the HexNAc oxonium ion as an indicator to target O-GlcNAc modified peptides with modification favorable ETD fragmentation, peptide sequencing and site localization can be achieved with improved selectivity and sensitivity.

Characterization of O-GlcNAc modified proteins enriched from HEK293T cell extract

186 proteins were identified at 1% pepFDR from the multiple-antibody enriched HEK293T cell extract. In comparison with the recently published work from our group[10] that used the same sample, 67% (124/186) of the identified proteins were consistent with the previous experiment, including heterogeneous nuclear ribonucleoproteins, BAT2 domain-containing protein 1, ribosomal proteins, heat shock proteins, and nuclear pore complex proteins. Among the rest 33% (62/186) of the dataset which had only been observed in our current study, 17 proteins were supported in literature as being O-GlcNAc modified and 45 were identified as novel O-GlcNAc proteins (Table 4-1). Furthermore, while the previous experiment did not discover any O-GlcNAc modification sites, our current study identified 58 HexNAc sites on 31 non-redundant peptides of 16 proteins. By comparing to literature, 16 of the 58 sites have been previously assigned leaving 42 novel O-GlcNAc sites (Table 4-2); and 15 of the 16 characterized proteins were also observed in the previously work from our group. In Figure 4-3, panel A and panel B are the consecutively acquired HCD and ETD spectra of the same triple-charged precursor m/z 760.3611. As noted in Figure 4-3A, most b- and y-

ions were generated without the attachment of HexNAc (b-HexNAc or y-HexNAc) and were at relatively low intensity, which were not useful in assigning the modification sites. A distinctive HexNAc oxonium ion peak was produced at m/z 204.09, and the fragment ions of HexNAc were also found at high intensities, indicating the presence of HexNAc modification on the precursor. In Figure 4-3B, an almost complete series of c- and z- ions were observed and the site of HexNAc modification was clearly indicated by corresponding c- and z- ion pairs. The two examples presented in Figure 4-3A-D confirmed the applicability of HCD/ETD approach to complex mixture samples, and further suggested the possibility of targeting HexNAc modified peptides with selective ETD fragmentation triggered by the HexNAc diagnostic ions detected under HCD.

We further analyzed the same raw spectra files obtained from the HEK293T sample using a multiple-engine database search approach to improve the sensitivity of data analysis (Figure 4-S1). Basically, database-search programs assign an MS^n spectrum with the most probable peptide match, and the assigned spectra generally constitute only a fraction of all the spectra acquired in an LC-MS^n analysis. Variations in database-searching algorithms for assigning peptides to MS^n spectra have been known to provide different identification results[27]. By combing search results from different search engines have been proven to lead to a larger number of protein identification with an increased rate of peptide assignments[28-30]. By processing the files in this manner, 200 proteins were identified at 1% pepFDR after combining and validating the output from three independent database searches. In comparison with the result from our recently published work[10] that used the same sample, 66% (132/200) of the identified proteins were consistent with the previous experiment, including heterogeneous nuclear

ribonucleoproteins, BAT2 domain-containing protein 1, ribosomal proteins, heat shock proteins, and nuclear pore complex proteins. Among the rest 34% (68/200) of the dataset which had only been observed in our current study, 18 proteins were supported in literature as being O-GlcNAc modified and 50 were identified as novel O-GlcNAc proteins. Furthermore, the multiple-engine search strategy indentified 165 HexNAc sites on 77 non-redundant peptides of 40 proteins, including the 58 sites that were assigned by SEQUEST-only database search. By comparing to literature, 29 of the 165 sites have been previously assigned leaving 136 novel O-GlcNAc sites; and 20 of the 40 characterized proteins were also observed in the previously work from our group[10], whereas 18 of the 40 proteins are novel. By taking advantage of the multiple-engine database search methodology, the numbers of identified proteins, peptides and modification sites have been greatly increased and the sensitivity of the acquired raw data has been dramatically improved.

OGT substrate specificity reflected by the microheterogeneity of O-GlcNAc modified peptides

O-GlcNAc modification is analogous to serine/threonine phosphorylation in many respects [15, 31]. However, unlike phosphorylation, which is catalyzed by almost 500 kinases encoded in human genome [32], O-GlcNAc modification is catalyzed by the products of a single human gene [33-34]. Studies have shown that OGT glycosylation is quite specific [35], and furthermore that the catalytic subunit of OGT achieves both high specificity and a remarkable diversity of substrates through forming a complex with a variety of targeting proteins via its tetratricopeptide repeat (TPR) protein-protein interaction domains[36-38]. In our dataset, we discovered several cases of clustered sites of

HexNAc modification exhibited by co-eluted peptides, including protein Nup214, Nup153, host cell factor and EMSY. As an example presented in Figure 4-4, multiple isobaric peptides were eluted at the same time and had acquired same charge state. In the same ETD scan, possibly three peptides with single HexNAc modifications on sequential serines (S7 in red, S8 in grey, and S9 in blue) contributed to the c- and z- ions observed in the spectrum. The peaks at m/z 969 and m/z 1056 correspond to the z10 and z11 ions of the S7-modified peptide; the peaks at m/z 1172 and m/z 1259 correspond to the z10 and z11 ions of the S9-modified peptide. All the ions observed in the spectrum can be explained as the fragments of S7-modified, S9-modified or both peptides, confirming the presence of both. However, since the peaks at m/z 969 and m/z 1259 correspond to the z10 and z11 ions of the S8-modified peptide and there is no unique ion for its elimination, it is possible that S8-modified peptide was also present in the co-eluted mixture. This observation indicated the existence of microheterogeneity in O-GlcNAc modification, and suggested the diversity of peptide substrates for OGT.

Diagnostic ion patterns observed from other types of glycosylated peptides under HCD

In order to explore the applicability of HCD/ETD MS scheme on other types of O-glycosylation, synthetic O-Mannose and O-GalNAc modified peptides were analyzed in the same fashion as O-GlcNAc modified peptides using the combined HCD/ETD fragmentation. Indicative ion patterns at low mass range were observed, respectively, when O-Mannose and O-GalNAc peptides underwent HCD: the O-GalNAc modified peptide exhibited similar signature ions at m/z 204, m/z 186, m/z 168, m/z 144, m/z 138, and m/z 126 (Figure 4-S2A and 4-S2B); the O-Mannosylated peptide exhibited the Hexose oxonium ion at m/z 163.05, and a series of signature ions at m/z 163, m/z 145,

m/z 127, m/z 115, and m/z 109 (Figure 4-S2A and 4-S2C, Table 4-S1). The utility of HCD has been demonstrated for the characterization of protein tyrosine-phosphorylation[20], protein N-glycosylation[39], and the quantification of iTRAQ-labeled phosphopeptides[40-41]. Our observation of the distinctive ion patterns of O-GlcNAc, O-Mannose, and O-GalNAc peptides proved the applicability of HCD in targeting O-glycosylated peptides. Although, based on our data, it appears that O-GlcNAc and O-GalNAc cannot be distinguished from each other solely by their signature ion patterns. Furthermore, the signature ion-trigger strategy has the potential to be applied to the analysis of post-translational modifications besides glycosylation and phosphorylation, such as methylation, bromylation, hydroxylation and other modifications that have previously been shown to produce specific fragment ions[42].

CONCLUSION

The diagnostic ion patterns of HexNAc and Hexose revealed under HCD condition can be utilized to selectively target HexNAc and/or Hexose modified proteins, especially when combined with ETD which preserves labile post-translational modification on proteins, the reliability and accuracy in glycoprotein identification and site localization can be greatly improved. In our study, we investigated the applicability of the combined HCD/ETD MS scheme in characterizing O-GlcNAc modified proteins from a complex biological sample and successfully identified 58 modification sites. Additionally, with a multiple-engine database search method, we were able to increase the sensitivity of our discovery drastically to reach a total of 165 sites of O-GlcNAc modification. We further proved the capability of the HCD/ETD scheme in characterizing O-GalNAc and/or O-Mannose modified peptides, and explored its potential in other

modified proteins. Along with the advancement of both hardware and software in mass spectrometry, we anticipate that an HCD-trigger-ETD approach will be implemented and realized on a hybrid linear ion trap/Orbitrap platform in the near future.

REFERENCES

1. Love, D. C.; Hanover, J. A., The hexosamine signaling pathway: deciphering the "O-GlcNAc code". *Sci STKE* **2005,** 2005, (312), re13.

2. Zachara, N. E.; Hart, G. W., O-GlcNAc a sensor of cellular state: the role of nucleocytoplasmic glycosylation in modulating cellular function in response to nutrition and stress. *Biochim Biophys Acta* **2004,** 1673, (1-2), 13-28.

3. Hart, G. W.; Housley, M. P.; Slawson, C., Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* **2007,** 446, (7139), 1017-22.

4. Laczy, B.; Hill, B. G.; Wang, K.; Paterson, A. J.; White, C. R.; Xing, D.; Chen, Y. F.; Darley-Usmar, V.; Oparil, S.; Chatham, J. C., Protein O-GlcNAcylation: a new signaling paradigm for the cardiovascular system. *Am J Physiol Heart Circ Physiol* **2009,** 296, (1), H13-28.

5. Copeland, R. J.; Bullen, J. W.; Hart, G. W., Cross-talk between GlcNAcylation and phosphorylation: roles in insulin resistance and glucose toxicity. *Am J Physiol Endocrinol Metab* **2008,** 295, (1), E17-28.

6. Dias, W. B.; Hart, G. W., O-GlcNAc modification in diabetes and Alzheimer's disease. *Mol Biosyst* **2007,** 3, (11), 766-72.

7. Lefebvre, T.; Guinez, C.; Dehennaut, V.; Beseme-Dekeyser, O.; Morelle, W.; Michalski, J. C., Does O-GlcNAc play a role in neurodegenerative diseases? *Expert Rev Proteomics* **2005,** 2, (2), 265-75.

8. Hu, P.; Shimoji, S.; Hart, G. W., Site-specific interplay between O-GlcNAcylation and phosphorylation in cellular regulation. *FEBS Lett* **2010,** 584, (12), 2526-38.

9. Haynes, P. A.; Aebersold, R., Simultaneous detection and identification of O-GlcNAc-modified glycoproteins using liquid chromatography-tandem mass spectrometry. *Anal Chem* **2000,** 72, (21), 5402-10.

10. Teo, C. F.; Ingale, S.; Wolfert, M. A.; Elsayed, G. A.; Not, L. G.; Chatham, J. C.; Wells, L.; Boons, G. J., Glycopeptide-specific monoclonal antibodies suggest new roles for O-GlcNAc. *Nat Chem Biol* **2010,** 6, (5), 338-43.

11. Comer, F. I.; Vosseller, K.; Wells, L.; Accavitti, M. A.; Hart, G. W., Characterization of a mouse monoclonal antibody specific for O-linked N-acetylglucosamine. *Anal Biochem* **2001,** 293, (2), 169-77.

12. Snow, C. M.; Senior, A.; Gerace, L., Monoclonal antibodies identify a group of nuclear pore complex glycoproteins. *J Cell Biol* **1987,** 104, (5), 1143-56.

13. Chalkley, R. J.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L., Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc Natl Acad Sci U S A* **2009,** 106, (22), 8894-9.

14. Vosseller, K.; Trinidad, J. C.; Chalkley, R. J.; Specht, C. G.; Thalhammer, A.; Lynn, A. J.; Snedecor, J. O.; Guan, S.; Medzihradszky, K. F.; Maltby, D. A.; Schoepfer, R.; Burlingame, A. L., O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. *Mol Cell Proteomics* **2006,** 5, (5), 923-34.

15. Wang, Z.; Udeshi, N. D.; Slawson, C.; Compton, P. D.; Sakabe, K.; Cheung, W. D.; Shabanowitz, J.; Hunt, D. F.; Hart, G. W., Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Sci Signal* **2010,** 3, (104), ra2.

16. Khidekel, N.; Ficarro, S. B.; Clark, P. M.; Bryan, M. C.; Swaney, D. L.; Rexach, J. E.; Sun, Y. E.; Coon, J. J.; Peters, E. C.; Hsieh-Wilson, L. C., Probing the dynamics of O-GlcNAc glycosylation in the brain using quantitative proteomics. *Nat Chem Biol* **2007,** 3, (6), 339-48.

17. Jebanathirajah, J.; Steen, H.; Roepstorff, P., Using optimized collision energies and high resolution, high accuracy fragment ion selection to improve glycopeptide detection by precursor ion scanning. *J Am Soc Mass Spectrom* **2003,** 14, (7), 777-84.

18. Chalkley, R. J.; Burlingame, A. L., Identification of GlcNAcylation sites of peptides and alpha-crystallin using Q-TOF mass spectrometry. *J Am Soc Mass Spectrom* **2001,** 12, (10), 1106-13.

19. Huddleston, M. J.; Bean, M. F.; Carr, S. A., Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests. *Anal Chem* **1993,** 65, (7), 877-84.

20. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **2007,** 4, (9), 709-12.

21. Carr, S. A.; Huddleston, M. J.; Bean, M. F., Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Sci* **1993,** 2, (2), 183-96.

22. Lim, J. M.; Sherling, D.; Teo, C. F.; Hausman, D. B.; Lin, D.; Wells, L., Defining the regulated secreted proteome of rodent adipocytes upon the induction of insulin resistance. *J Proteome Res* **2008,** 7, (3), 1251-63.

23. Scherl, A.; Shaffer, S. A.; Taylor, G. K.; Hernandez, P.; Appel, R. D.; Binz, P. A.; Goodlett, D. R., On the benefits of acquiring peptide fragment ions at high measured mass accuracy. *J Am Soc Mass Spectrom* **2008,** 19, (6), 891-901.

24. Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E.; Shabanowitz, J.; Hunt, D. F., The utility of ETD mass spectrometry in proteomic analysis. *Biochim Biophys Acta* **2006,** 1764, (12), 1811-22.

25. Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E.; Coon, J. J., Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Anal Chem* **2007,** 79, (2), 477-85.

26. Peterman, S. M.; Mulholland, J. J., A novel approach for identification and characterization of glycoproteins using a hybrid linear ion trap/FT-ICR mass spectrometer. *J Am Soc Mass Spectrom* **2006,** 17, (2), 168-79.

27. Boutilier, K.; Ross, M.; Podtelejnikov, A. V.; Orsi, C.; Taylor, R.; Taylor, P.; Figeys, D., Comparison of different search engines using validated MS/MS test datasets. *Analytica Chimica Acta* **2005,** 534, 10.

28. Yu, W.; Taylor, J. A.; Davis, M. T.; Bonilla, L. E.; Lee, K. A.; Auger, P. L.; Farnsworth, C. C.; Welcher, A. A.; Patterson, S. D., Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. *Proteomics* **2010,** 10, (6), 1172-89.

29. Carrascal, M.; Gay, M.; Ovelleiro, D.; Casas, V.; Gelpi, E.; Abian, J., Characterization of the human plasma phosphoproteome using linear ion trap mass spectrometry and multiple search engines. *J Proteome Res* **2010,** 9, (2), 876-84.

30. Searle, B. C.; Turner, M.; Nesvizhskii, A. I., Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res* **2008,** 7, (1), 245-53.

31. Wang, Z.; Gucek, M.; Hart, G. W., Cross-talk between GlcNAcylation and phosphorylation: site-specific phosphorylation dynamics in response to globally elevated O-GlcNAc. *Proc Natl Acad Sci U S A* **2008,** 105, (37), 13793-8.

32. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S., The protein kinase complement of the human genome. *Science* **2002,** 298, (5600), 1912-34.

33. Nolte, D.; Muller, U., Human O-GlcNAc transferase (OGT): genomic structure, analysis of splice variants, fine mapping in Xq13.1. *Mamm Genome* **2002,** 13, (1), 62-4.

34. Shafi, R.; Iyer, S. P.; Ellies, L. G.; O'Donnell, N.; Marek, K. W.; Chui, D.; Hart, G. W.; Marth, J. D., The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc Natl Acad Sci U S A* **2000,** 97, (11), 5735-9.

35. Lubas, W. A.; Hanover, J. A., Functional expression of O-linked GlcNAc transferase. Domain structure and substrate specificity. *J Biol Chem* **2000,** 275, (15), 10983-8.

36. Cheung, W. D.; Sakabe, K.; Housley, M. P.; Dias, W. B.; Hart, G. W., O-linked beta-N-acetylglucosaminyltransferase substrate specificity is regulated by myosin phosphatase targeting and other interacting proteins. *J Biol Chem* **2008,** 283, (49), 33935-41.

37. Iyer, S. P.; Hart, G. W., Roles of the tetratricopeptide repeat domain in O-GlcNAc transferase targeting and protein substrate specificity. *J Biol Chem* **2003,** 278, (27), 24608-16.

38. Kreppel, L. K.; Hart, G. W., Regulation of a cytosolic and nuclear O-GlcNAc transferase. Role of the tetratricopeptide repeats. *J Biol Chem* **1999,** 274, (45), 32015-22.

39. Segu, Z. M.; Mechref, Y., Characterizing protein glycosylation sites through higher-energy C-trap dissociation. *Rapid Commun Mass Spectrom* **2010,** 24, (9), 1217-25.

40. Zhang, Y.; Ficarro, S. B.; Li, S.; Marto, J. A., Optimized Orbitrap HCD for quantitative analysis of phosphopeptides. *J Am Soc Mass Spectrom* **2009,** 20, (8), 1425-34.

41. Boja, E. S.; Phillips, D.; French, S. A.; Harris, R. A.; Balaban, R. S., Quantitative mitochondrial phosphoproteomics using iTRAQ on an LTQ-Orbitrap with high energy collision dissociation. *J Proteome Res* **2009,** 8, (10), 4665-75.

42. Carr, S. A.; Annan, R. S.; Huddleston, M. J., Mapping posttranslational modifications of proteins by MS-based selective detection: application to phosphoproteomics. *Methods Enzymol* **2005,** 405, 82-115.

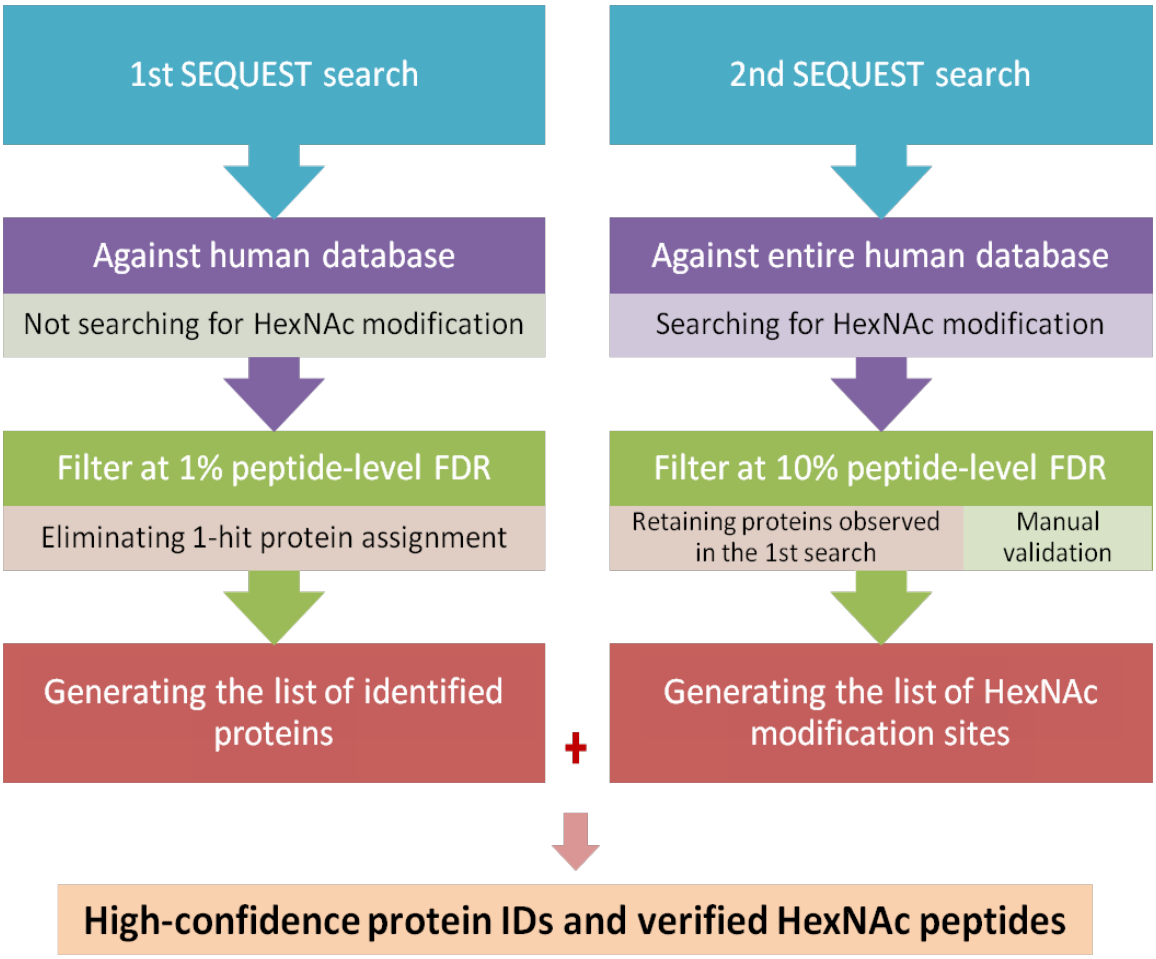
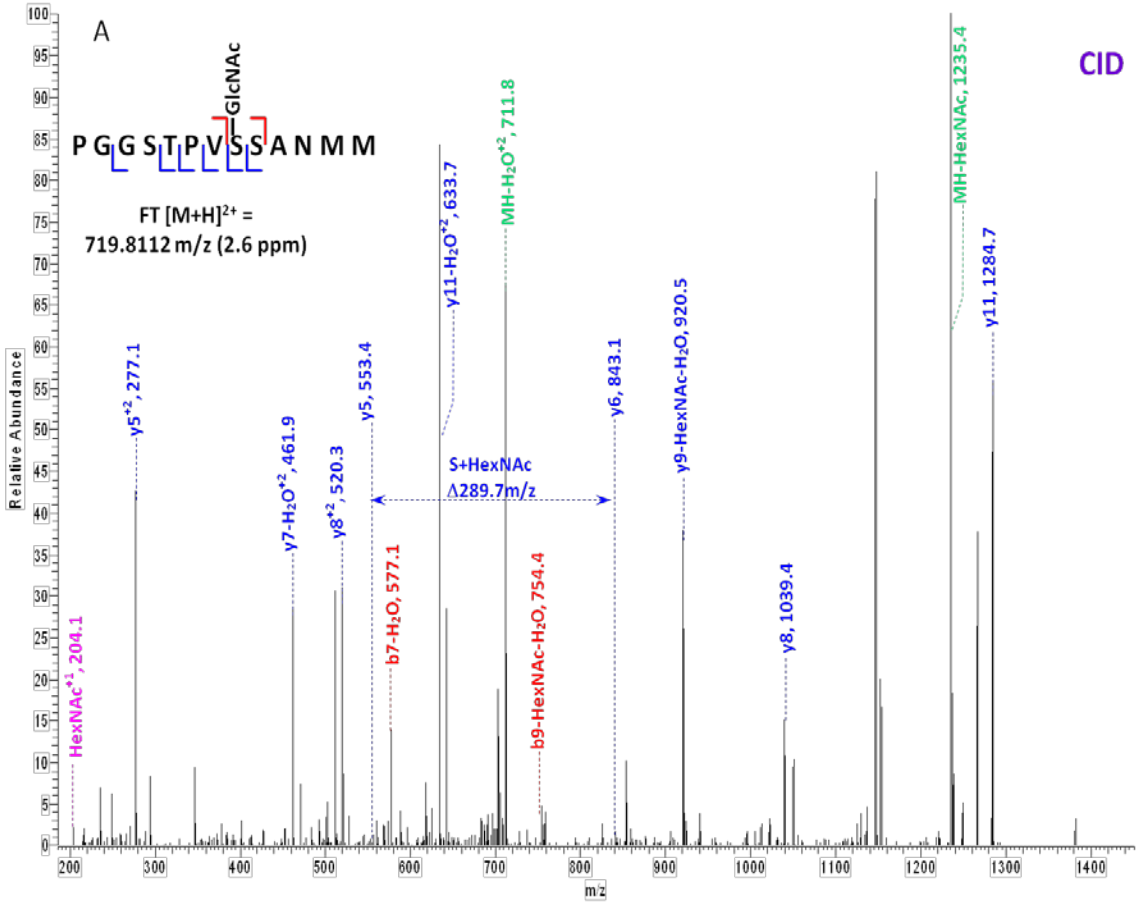
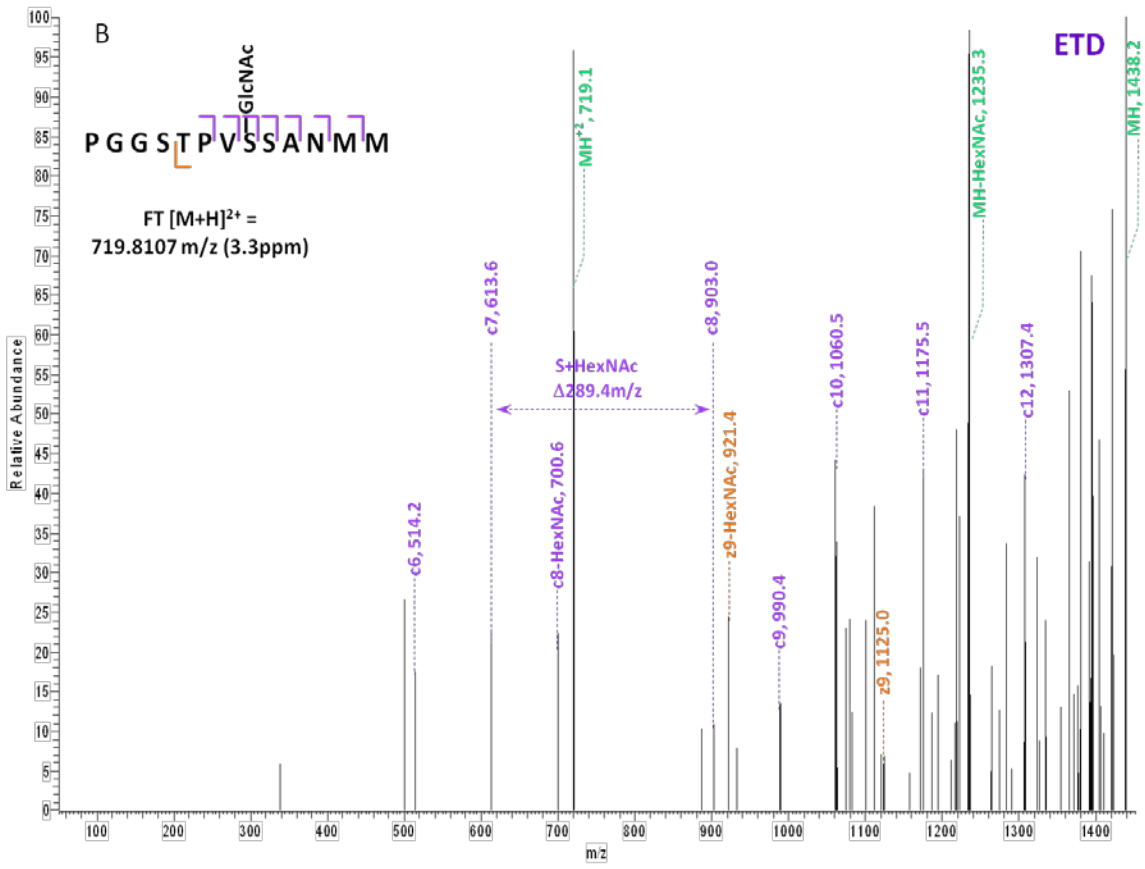
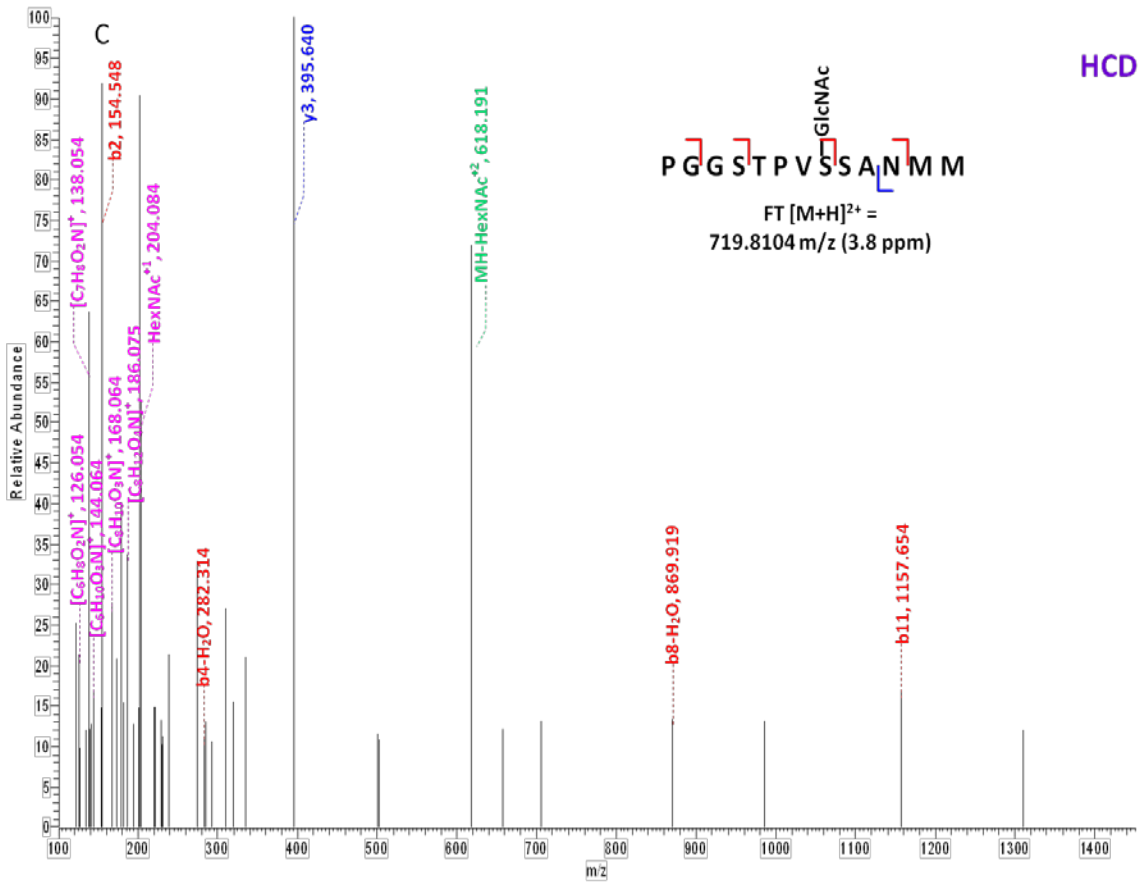
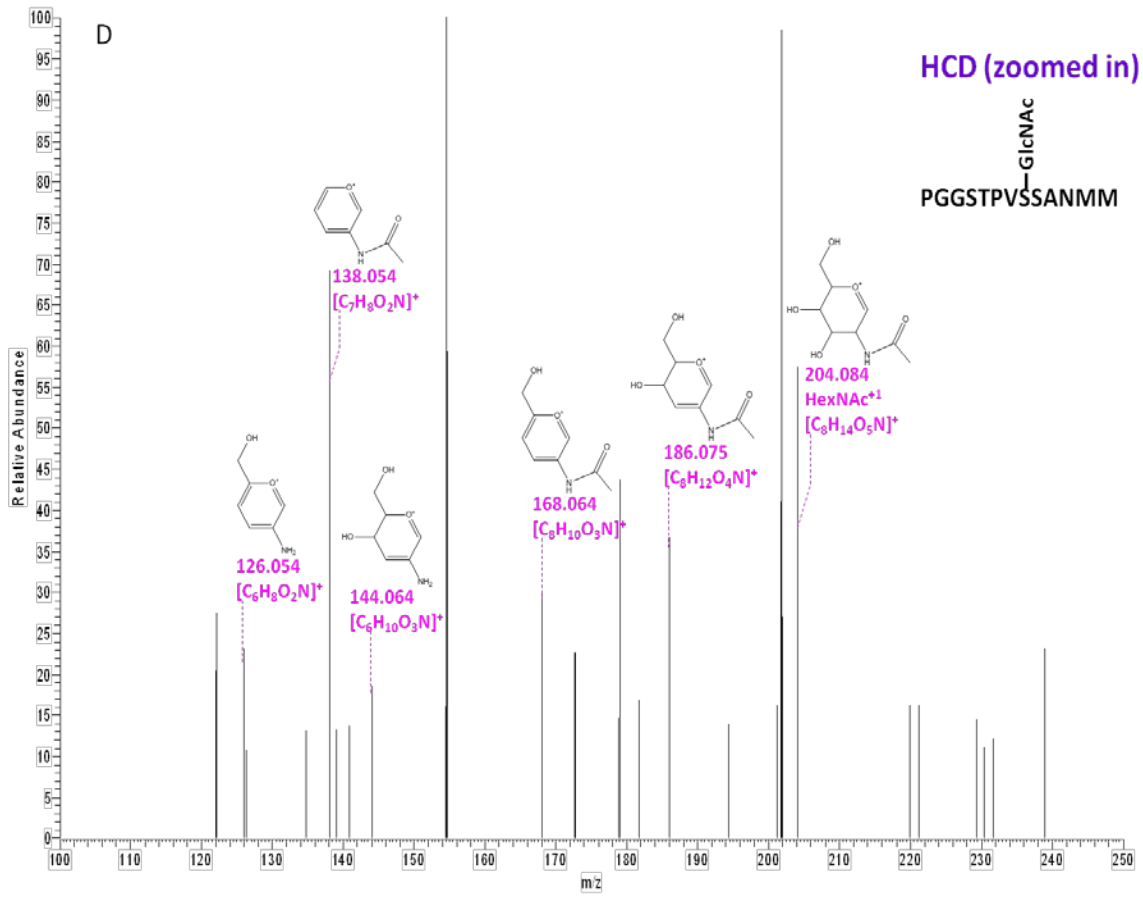


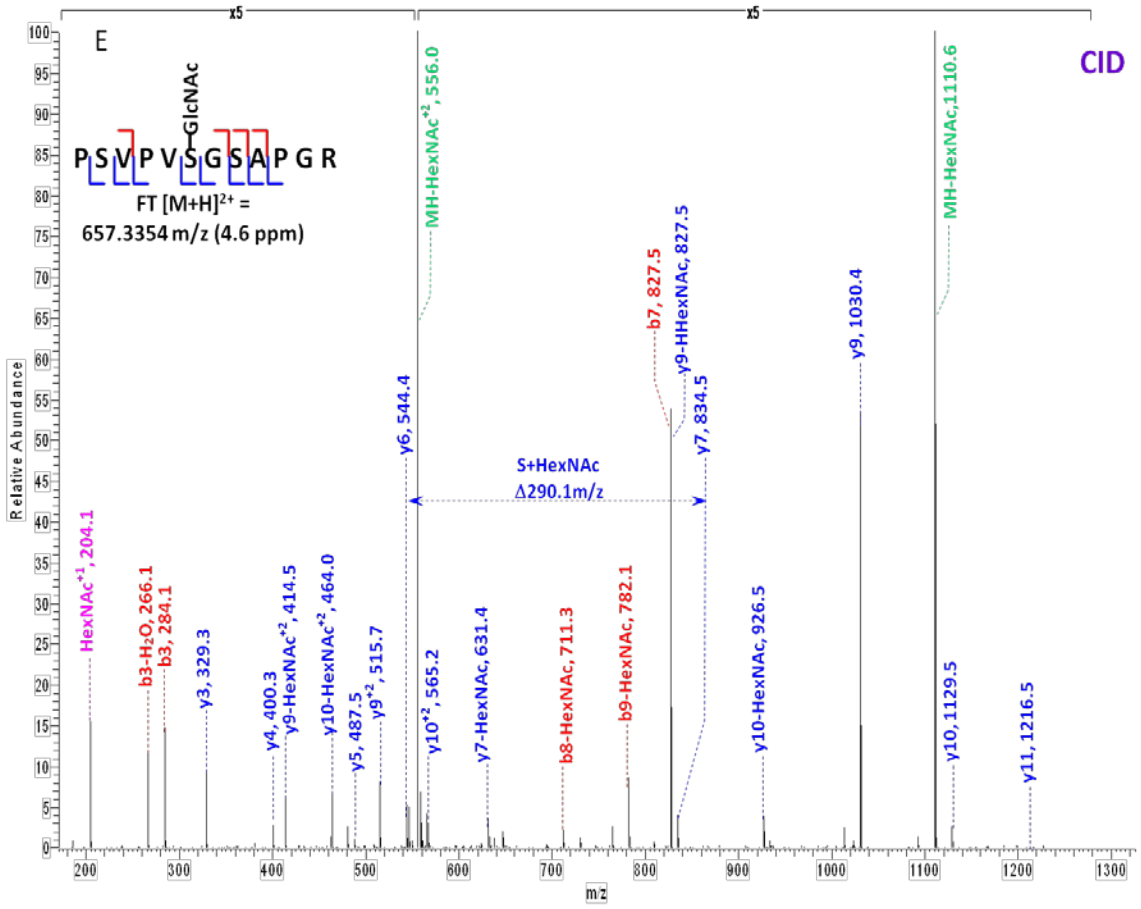
Figure 4-1. Database search strategy for the enriched HEK293T sample.
Raw spectra were searched twice against human database using SEQUEST. The output of the first search was generated without allowing for HexNAc modification and produced the list of identified proteins after filtering at 1% peptide-level FDR and eliminating 1-hit protein assignments; the output of the second search was generated allowing for HexNAc modification and produced the list of GlcNAc modified peptides after filtering at 10% peptide-level FDR, retaining proteins identified in the first search, and manual validation of the spectra.

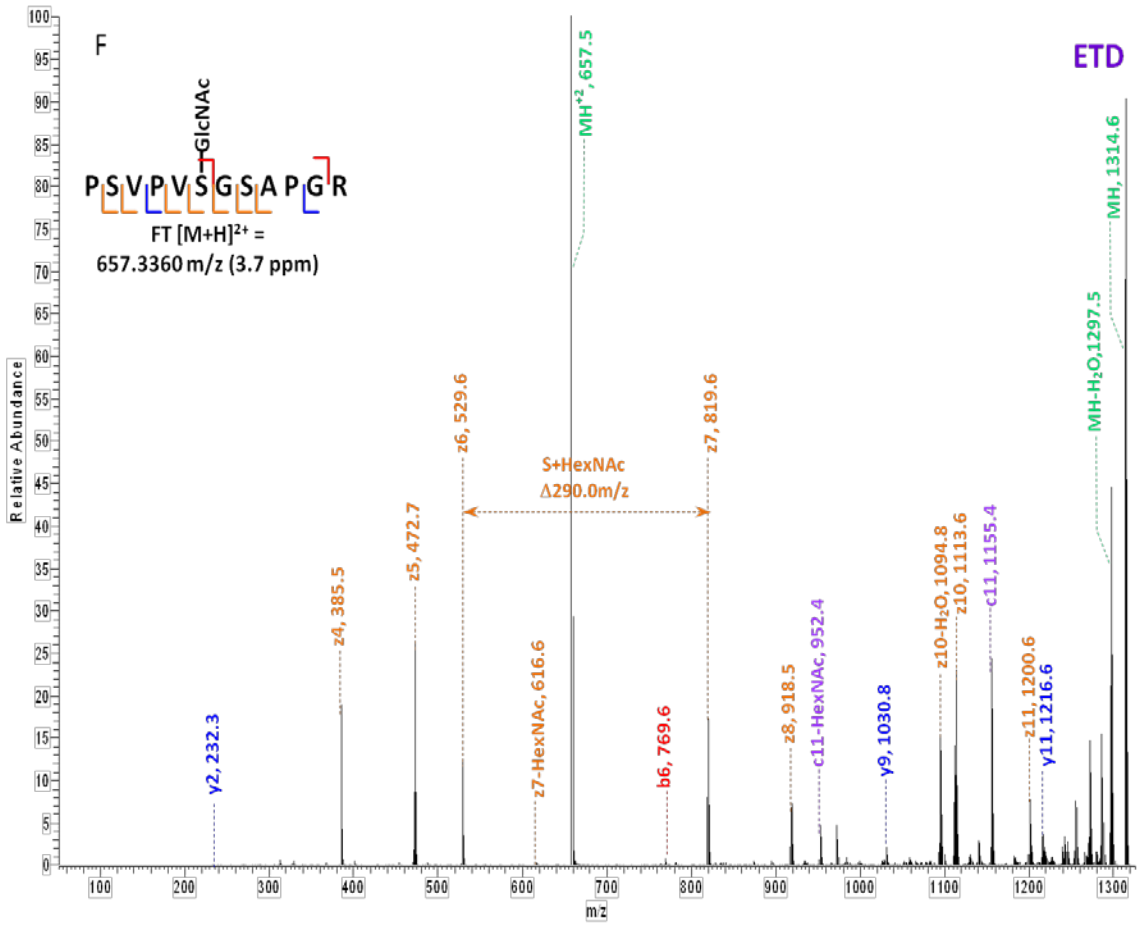

A
GlcNAc
P G G S T P V S S A N M M
FT [M+H]2+ =
719.8112 m/z (2.6 ppm)
CID
Relative Abundance
m/z
HexNAc+1, 204.1
y5+2, 277.1
y7-H2O+2, 461.9
y8+2, 520.3
y5, 553.4
b7-H2O, 577.1
y11-H2O+2, 633.7
MH-H2O+2, 711.8
S+HexNAc
Δ289.7m/z
b9-HexNAc-H2O, 754.4
y6, 843.1
y9-HexNAc-H2O, 920.5
y8, 1039.4
MH-HexNAc, 1235.4
y11, 1284.7

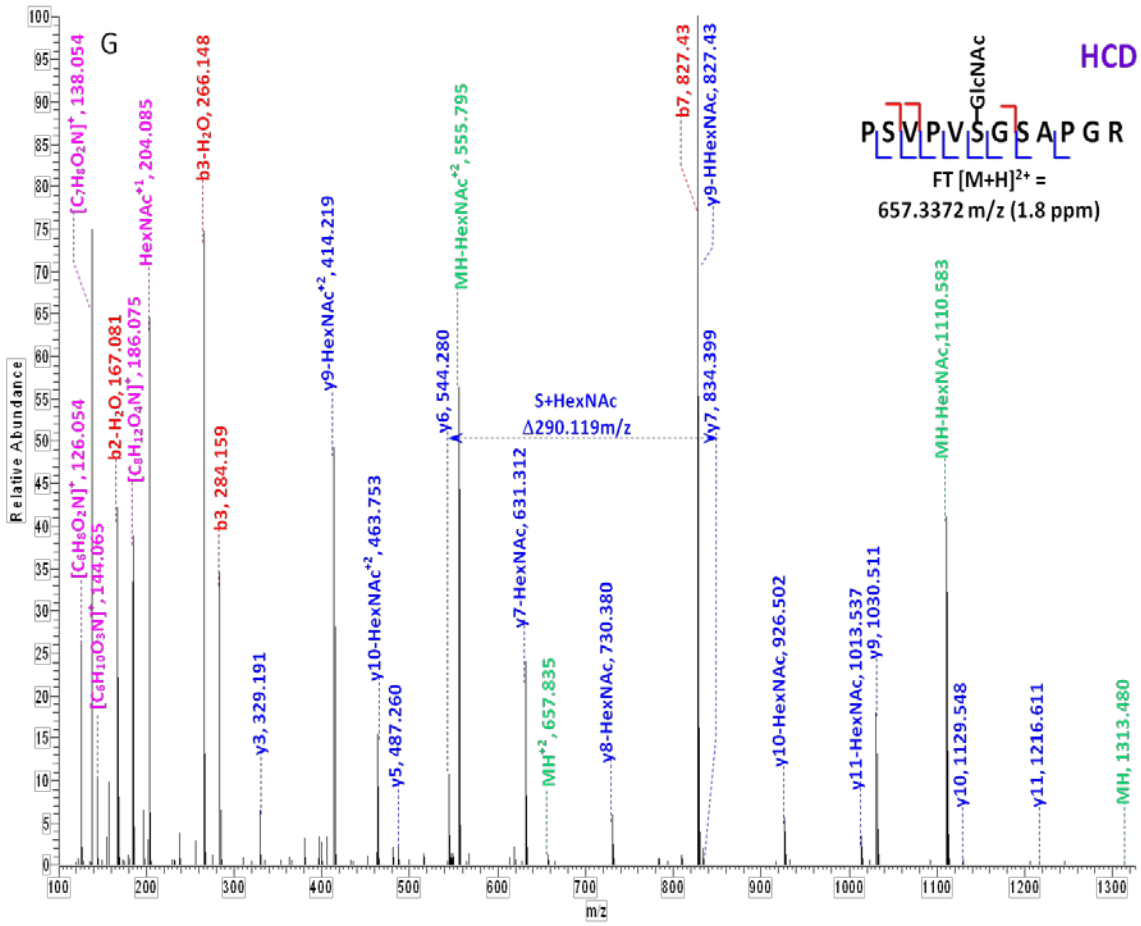

B
GlcNAc
P G G S T P V S S A N M M
FT [M+H]2+ =
719.8107 m/z (3.3ppm)
ETD
MH+2, 719.1
MH, 1438.2
MH-HexNAc, 1235.3
c7, 613.6
c8, 903.0
c10, 1060.5
c11, 1175.5
c12, 1307.4
S+HexNAc
Δ289.4m/z
c8-HexNAc, 700.6
z9-HexNAc, 921.4
c6, 514.2
c9, 990.4
z9, 1125.0
Relative Abundance
m/z


C
HCD
GlcNAc
P G G S T P V S S A N M M
FT [M+H]2+ =
719.8104 m/z (3.8 ppm)
Relative Abundance
m/z
y3, 395.640
b2, 154.548
[C7H8O2N]+, 138.054
MH-HexNAc*2, 618.191
HexNAc*1, 204.084
[C6H8O2N]+, 126.054
[C6H10O2N]+, 144.064
[C6H10O2N]+, 168.064
[C6H12O2N]+, 186.075
b4-H2O, 282.314
b8-H2O, 869.919
b11, 1157.654


D
HCD (zoomed in)
GlcNAc
PGGSTPVSSANMM
Relative Abundance
m/z
126.054
[C6H8O2N]+
138.054
[C7H8O2N]+
144.064
[C6H10O3N]+
168.064
[C8H10O3N]+
186.075
[C8H12O4N]+
204.084
HexNAc+1
[C8H14O5N]+


E
PSVPVSGSAPGR
GlcNAc
FT [M+H]2+ =
657.3354 m/z (4.6 ppm)
x5
x5
CID
Relative Abundance
m/z
MH-HexNAc+2, 556.0
MH-HexNAc,1110.6
b7, 827.5
y9-HHexNAc, 827.5
y7, 834.5
y9, 1030.4
y6, 544.4
S+HexNAc
Δ290.1m/z
HexNAc+1, 204.1
b3-H2O, 266.1
b3, 284.1
y3, 329.3
y4, 400.3
y9-HexNAc+2, 414.5
y10-HexNAc+2, 464.0
y5, 487.5
y9+2, 515.7
y10+2, 565.2
y7-HexNAc, 631.4
b8-HexNAc, 711.3
b9-HexNAc, 782.1
y10-HexNAc, 926.5
y10, 1129.5
y11, 1216.5


F
GlcNAc
PSVPVSGSAPGR
FT [M+H]2+ =
657.3360 m/z (3.7 ppm)
ETD
MH+2, 657.5
MH, 1314.6
MH-H2O,1297.5
z6, 529.6
z7, 819.6
S+HexNAc
Δ290.0m/z
z5, 472.7
z4, 385.5
z7-HexNAc, 616.6
b6, 769.6
z8, 918.5
c11-HexNAc, 952.4
y9, 1030.8
z10-H2O,1094.8
z10, 1113.6
c11, 1155.4
z11, 1200.6
y11, 1216.6
y2, 232.3
Relative Abundance
m/z


HCD
PSVPVSGSAPGR
GlcNAc
FT [M+H]2+ =
657.3372 m/z (1.8 ppm)
Relative Abundance
m/z
S+HexNAc
Δ290.119m/z
[C7H8O2N]+, 138.054
[C6H8O2N]+, 126.054
[C6H10O3N]+, 144.065
b2-H2O, 167.081
[C8H12O4N]+, 186.075
HexNAc+1, 204.085
b3-H2O, 266.148
b3, 284.159
y3, 329.191
y9-HexNAc+2, 414.219
y10-HexNAc+2, 463.753
y5, 487.260
MH-HexNAc+2, 555.795
y6, 544.280
y7-HexNAc, 631.312
MH+2, 657.835
y8-HexNAc, 730.380
b7, 827.43
y9-HHexNAc, 827.43
y7, 834.399
y10-HexNAc, 926.502
y11-HexNAc, 1013.537
y9, 1030.511
MH-HexNAc, 1110.583
y10, 1129.548
y11, 1216.611
MH, 1313.480

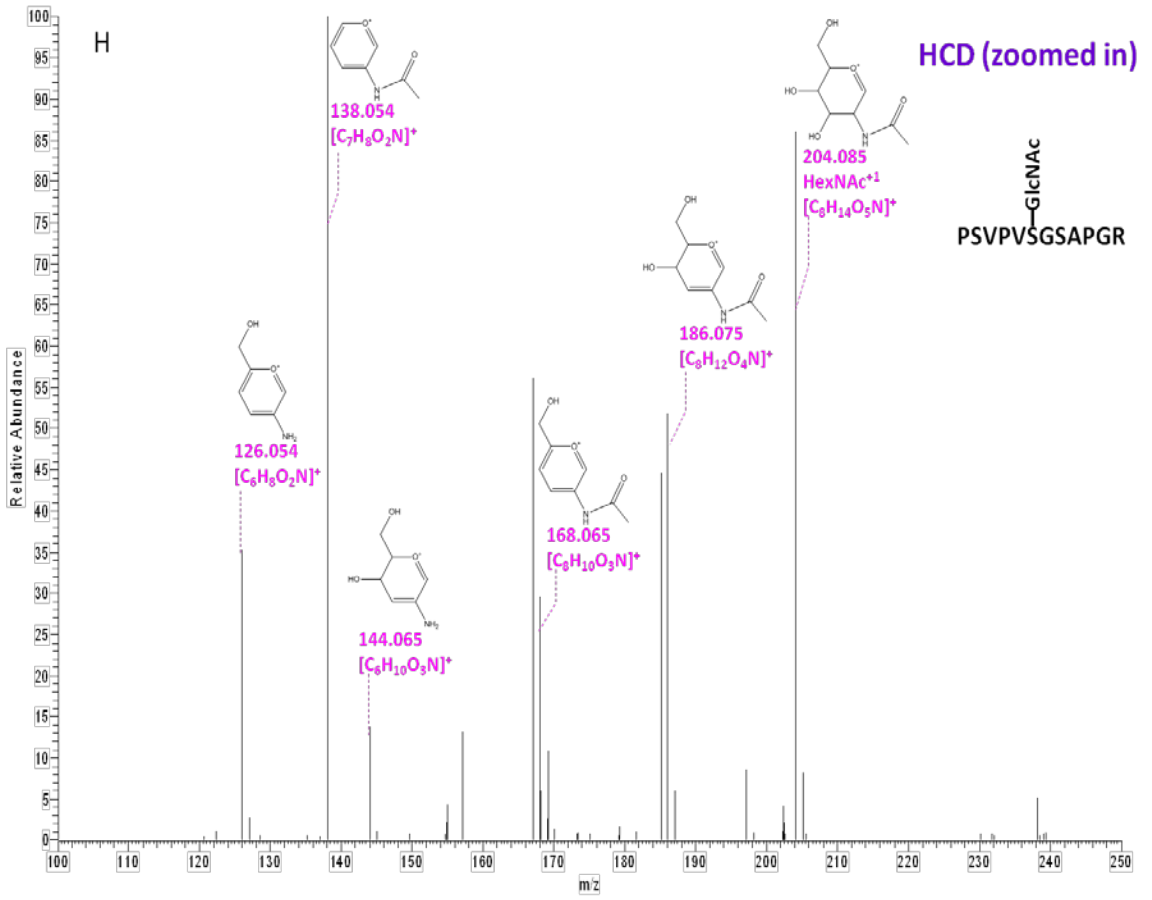
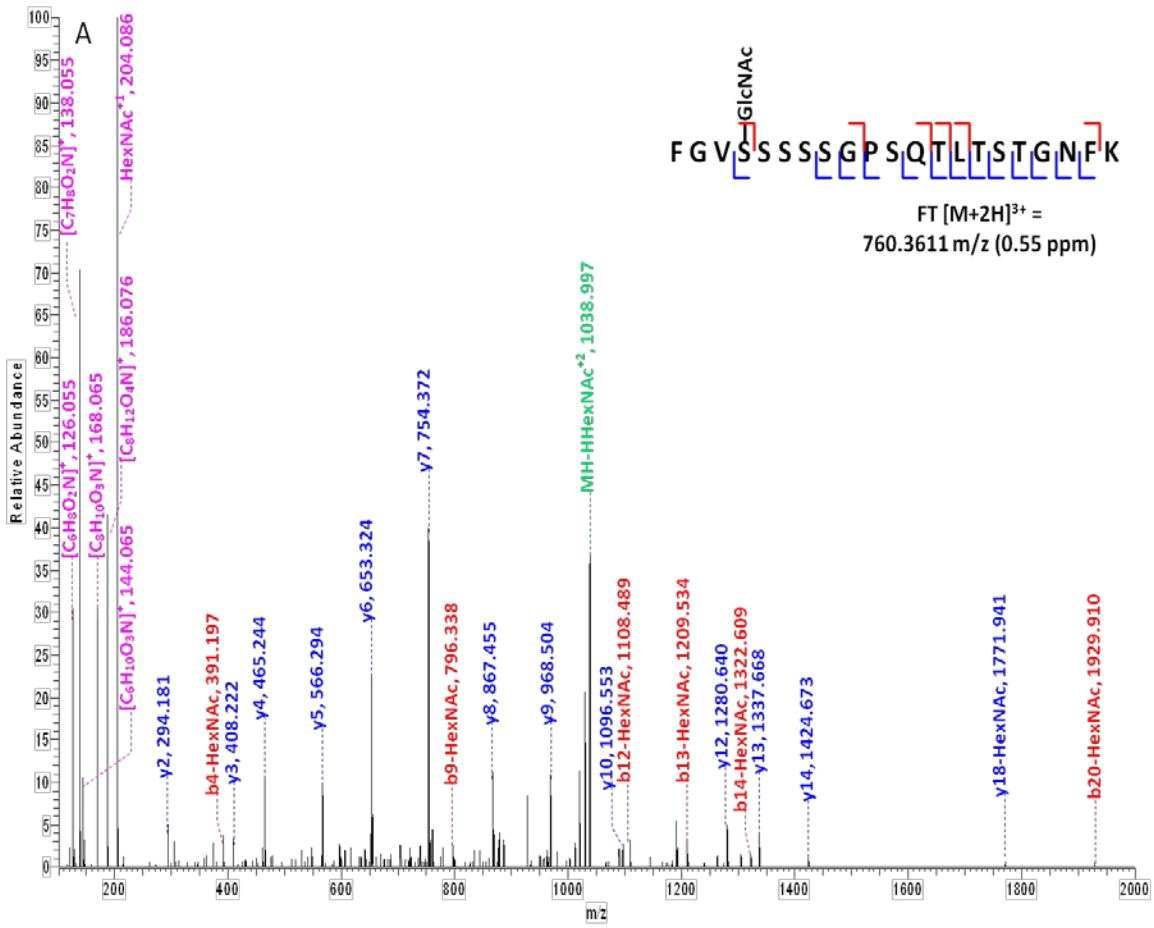
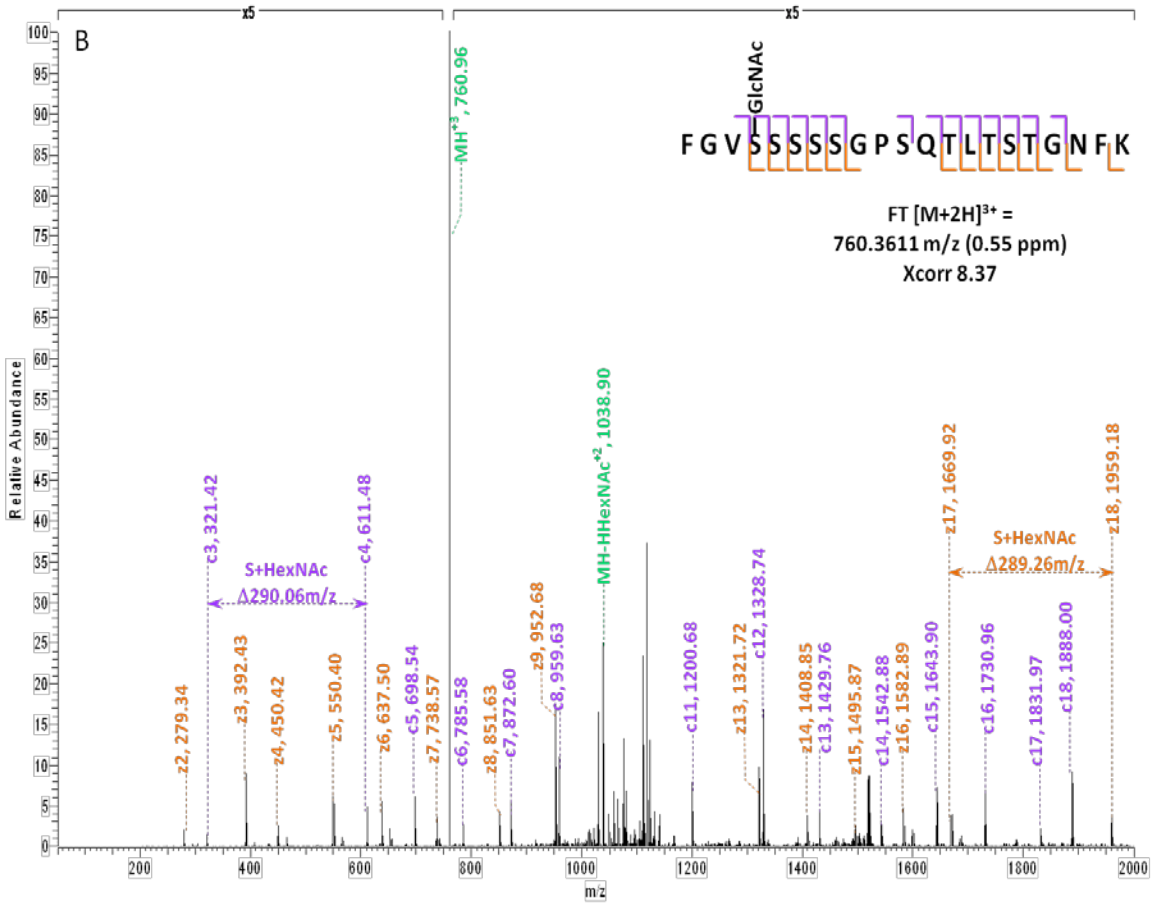
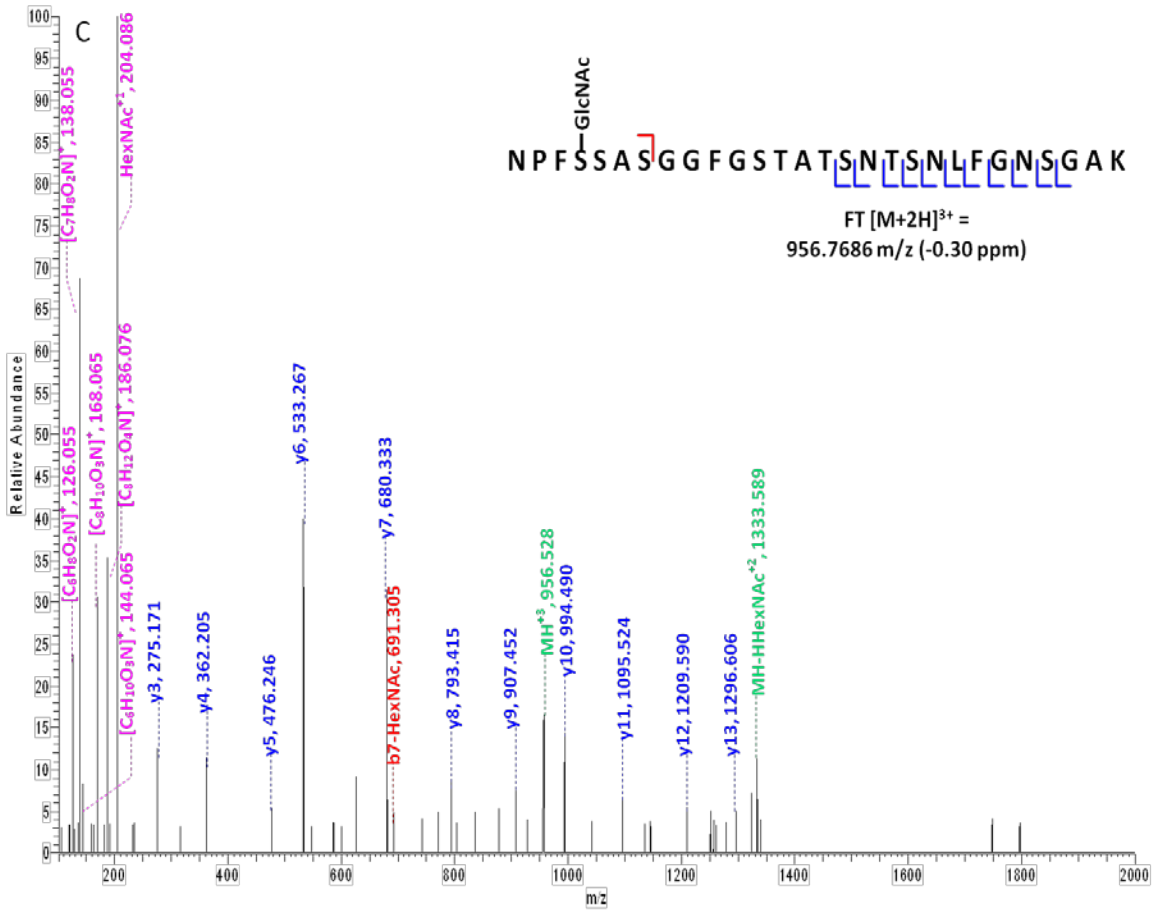


Figure 4-2. Respective CID, ETD, and HCD spectra of standard O-GlcNAc modified peptides CKII and BPP.

(**A**)-(**B**) CID and ETD spectra of O-GlcNAc modified CKII peptide, respectively; (**C**)-(**D**) HCD and zoomed in HCD spectra of O-GlcNAc modified CKII peptide; (**E**)-(**F**) CID and ETD spectra of O-GlcNAc modified BPP peptide, respectively; (**G**)-(**H**) HCD and zoomed in HCD spectra of O-GlcNAc modified CKII peptide. Note: "-HexNAc" or "-H2O" indicates the loss of HexNAc or H2O. The CID spectra of both peptides show HexNAc loss on the precursors and most b- and y- ions, and do not provide enough information for modification site localization. The ETD spectra show no HexNAc loss on the majority of c- and z- ions and provide intense ions for reliable site localization. The HCD spectra present similar HexNAc loss on most ions as the CID spectra, and at the low mass range a distinctive pattern of HexNAc fragments.


A
GlcNAc
F G V S S S S G P S Q T L T S T G N F K
FT [M+2H]3+ =
760.3611 m/z (0.55 ppm)
Relative Abundance
m/z
[C7H8O2N]+, 138.055
[C6H8O2N]+, 126.055
[C8H10O3N]+, 168.065
[C6H10O3N]+, 144.065
[C6H12O4N]+, 186.076
HexNAc+1, 204.086
y2, 294.181
b4-HexNAc, 391.197
y3, 408.222
y4, 465.244
y5, 566.294
y6, 653.324
y7, 754.372
b9-HexNAc, 796.338
y8, 867.455
y9, 968.504
MH-HHexNAc+2, 1038.997
y10, 1096.553
b12-HexNAc, 1108.489
b13-HexNAc, 1209.534
y12, 1280.640
b14-HexNAc, 1322.609
y13, 1337.668
y14, 1424.673
y18-HexNAc, 1771.941
b20-HexNAc, 1929.910


B
x5
x5
GlcNAc
FGVSSSSSGPSQTLTSTGNFK
FT [M+2H]3+ =
760.3611 m/z (0.55 ppm)
Xcorr 8.37
MH+3, 760.96
MH-HHexNAc+2, 1038.90
c3, 321.42
S+HexNAc
Δ290.06m/z
c4, 611.48
z2, 279.34
z3, 392.43
z4, 450.42
z5, 550.40
z6, 637.50
c5, 698.54
z7, 738.57
c6, 785.58
z8, 851.63
c7, 872.60
z9, 952.68
c8, 959.63
c11, 1200.68
z13, 1321.72
c12, 1328.74
z14, 1408.85
c13, 1429.76
z15, 1495.87
c14, 1542.88
z16, 1582.89
c15, 1643.90
z17, 1669.92
S+HexNAc
Δ289.26m/z
c16, 1730.96
c17, 1831.97
c18, 1888.00
z18, 1959.18
Relative Abundance
m/z


C
GlcNAc
NPFSSASGGFGSTATSNTSNLFGNSGAK
FT [M+2H]3+ =
956.7686 m/z (-0.30 ppm)
Relative Abundance
m/z
[C7H5O2N]+, 138.055
HexNAc+1, 204.086
[C6H8O2N]+, 126.055
[C8H10O3N]+, 168.065
[C8H12O4N]+, 186.076
[C6H10O3N]+, 144.065
y3, 275.171
y4, 362.205
y5, 476.246
y6, 533.267
b7-HexNAc, 691.305
y7, 680.333
y8, 793.415
y9, 907.452
MH+3, 956.528
y10, 994.490
y11, 1095.524
y12, 1209.590
y13, 1296.606
MH-HHexNAc+2, 1333.589

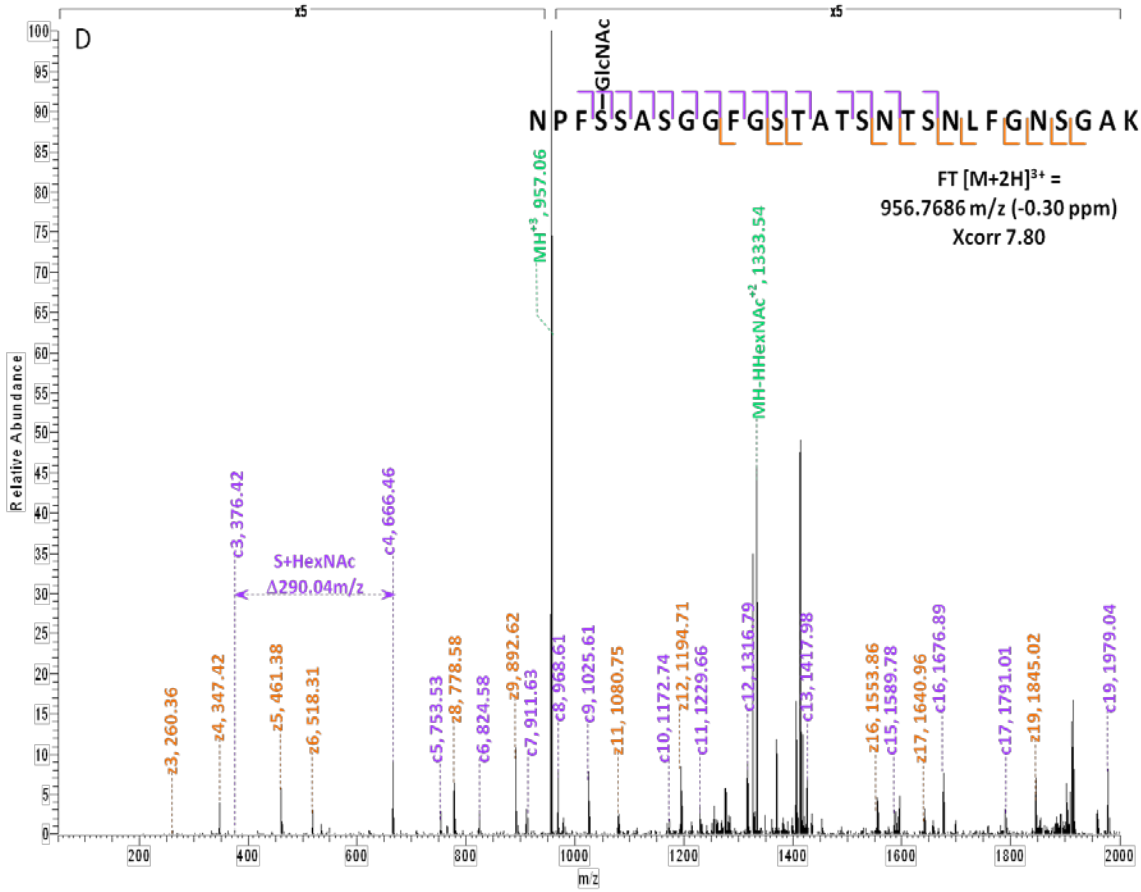


Figure 4-3. Corresponding HCD and ETD spectra of O-GlcNAc modified peptides identified in the enriched HEK293T sample.

(**A-B**) HCD and corresponding ETD spectra of peptide FGVS(HexNAc)SSSSGPSQTLTSTGNFK; (**C-D**) HCD and corresponding ETD spectra of peptide NPFS(HexNAc)SASGGFGSTATSNTSNLFGNSGAK. Note: "-HexNAc" or "-H2O" indicates the loss of HexNAc or H2O. The HCD spectra of both peptides exhibt the diagnostic ion pattern of HexNAc fragments at the low mass range, and both ETD spectra provide abundant indicative c- and z- ions for HexNAc modification site assignment.

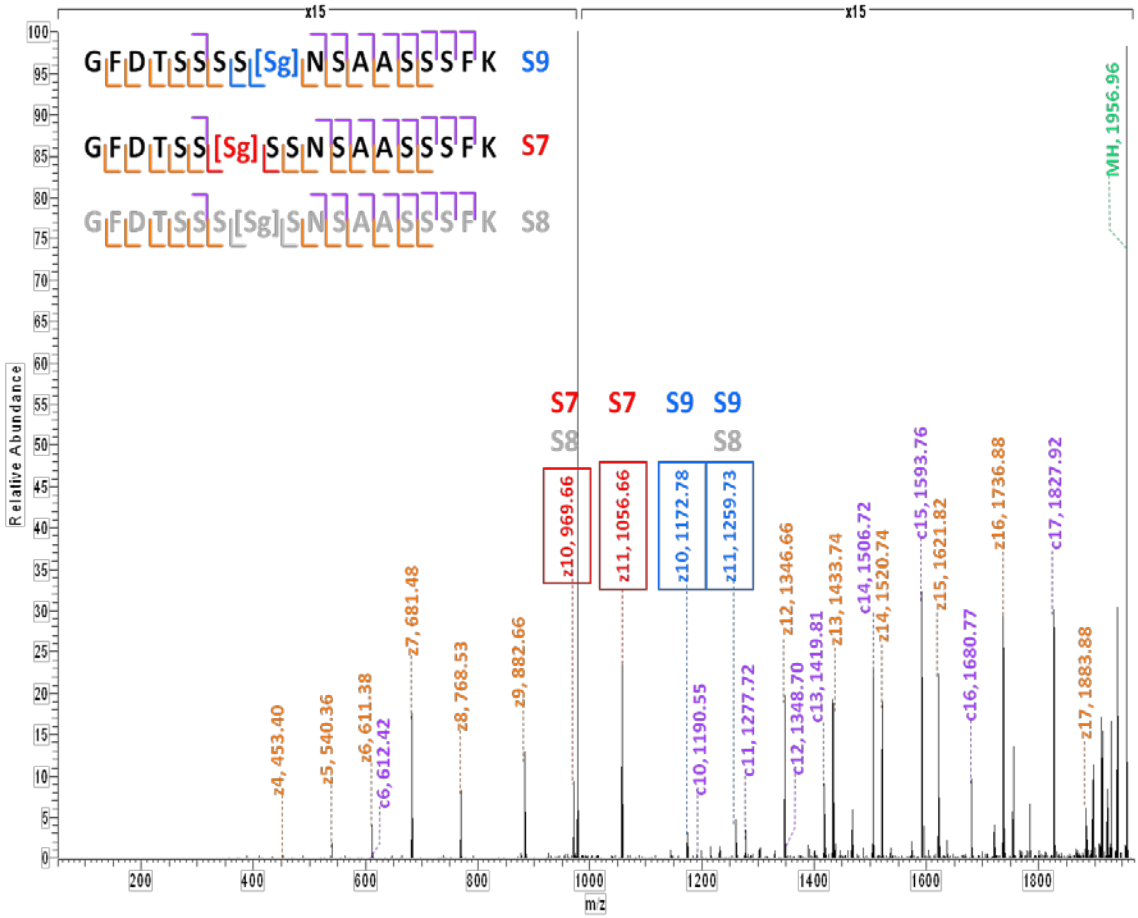


Figure 4-4. ETD spectrum of co-eluted O-GlcNAc modified peptides GFDTSSSSs(HexNAc)NSAASSSFK and GFDTSSs(HexNAc)SSNSAASSSFK.

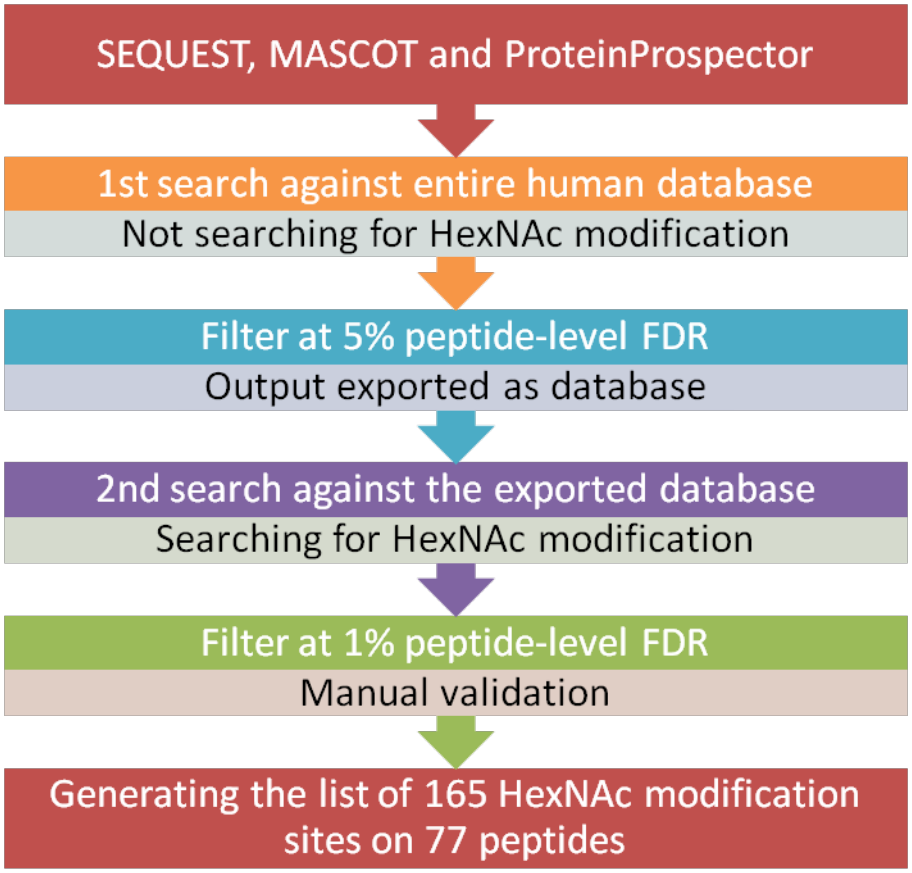
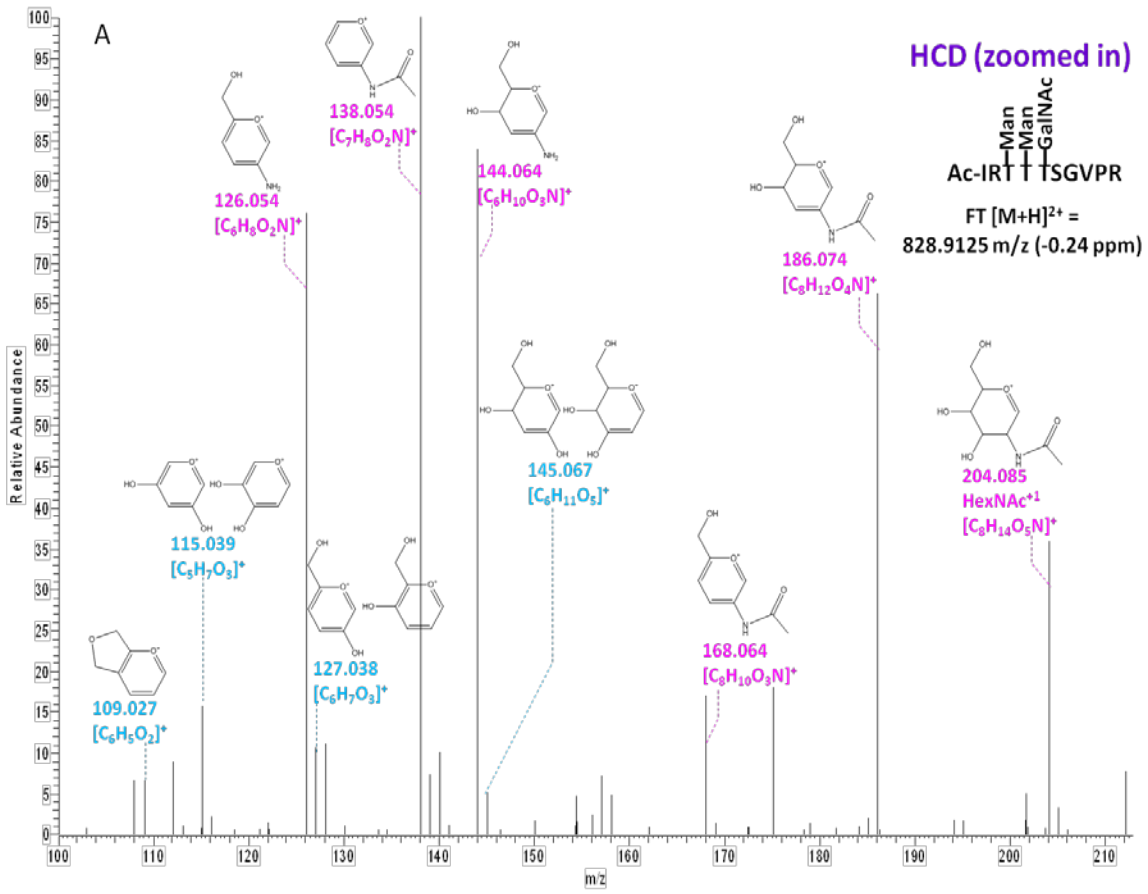
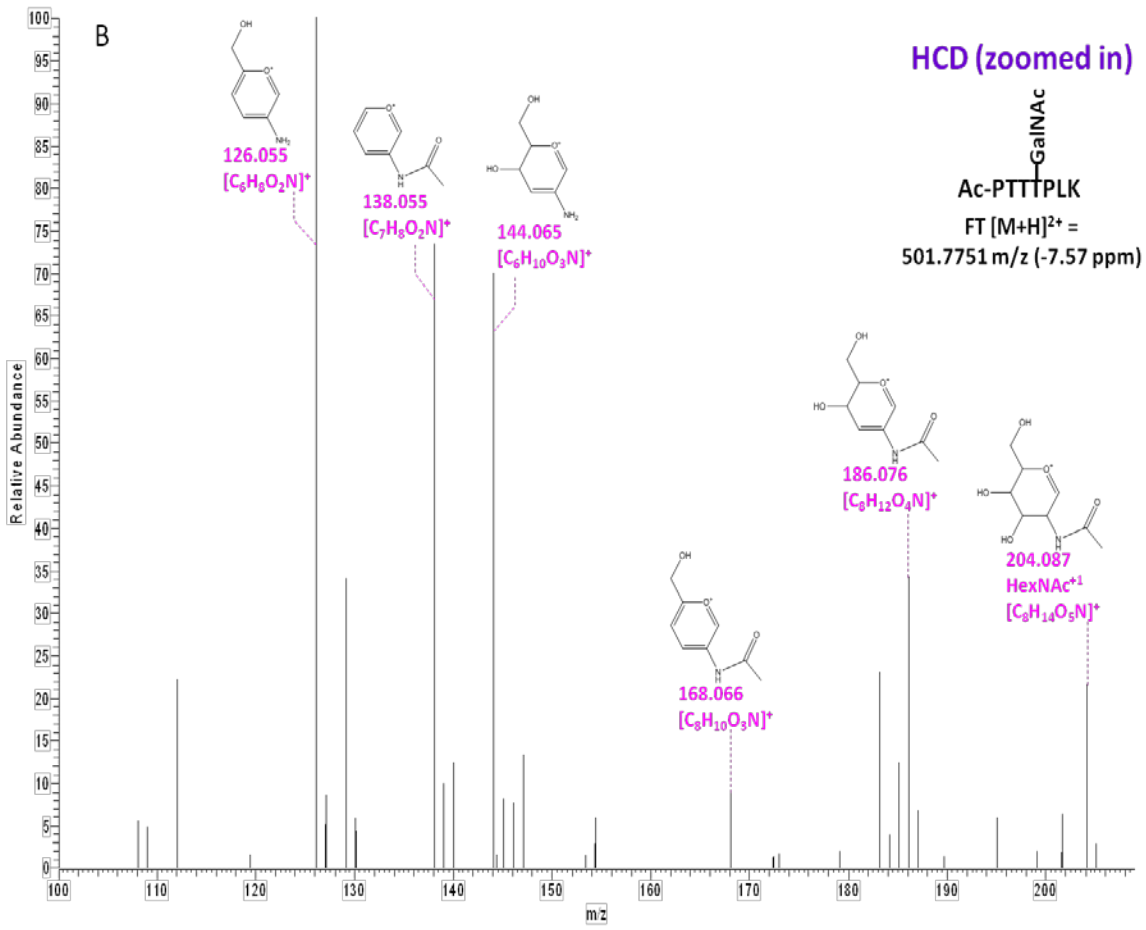
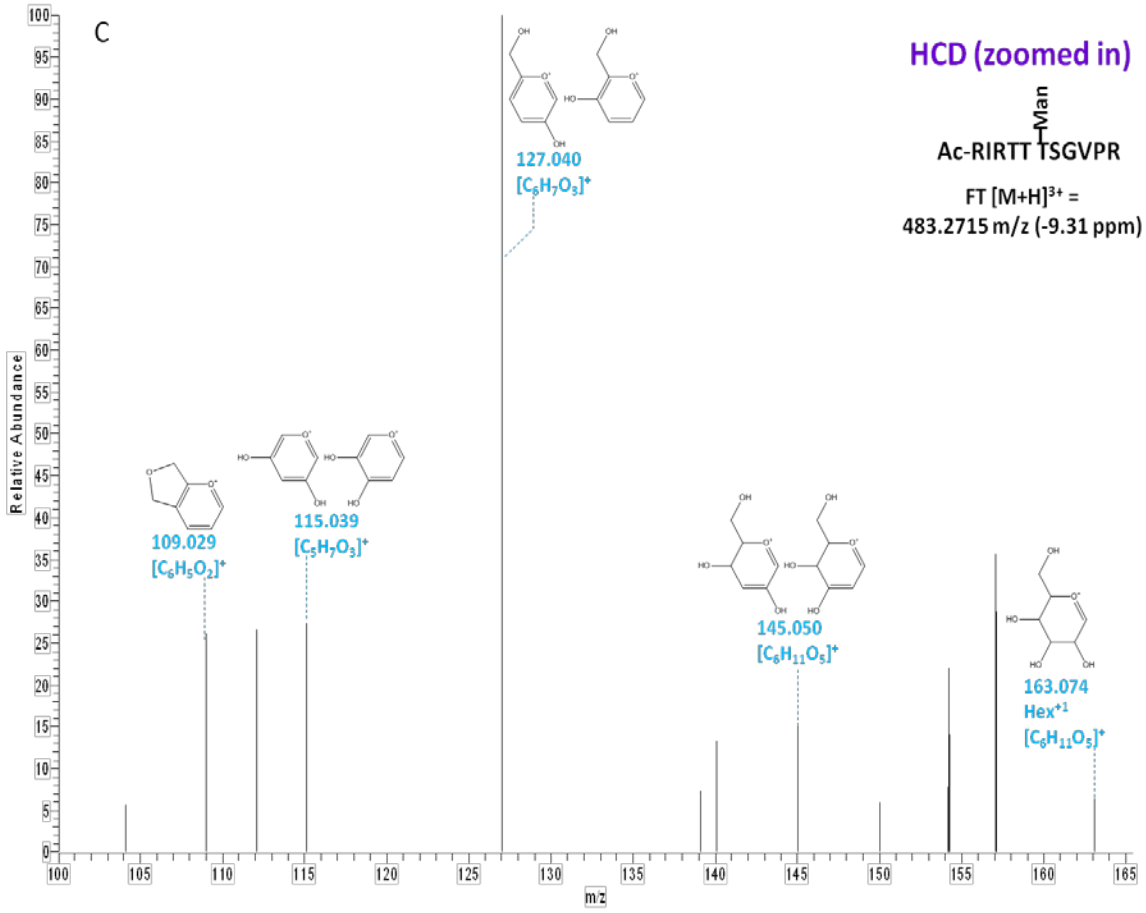
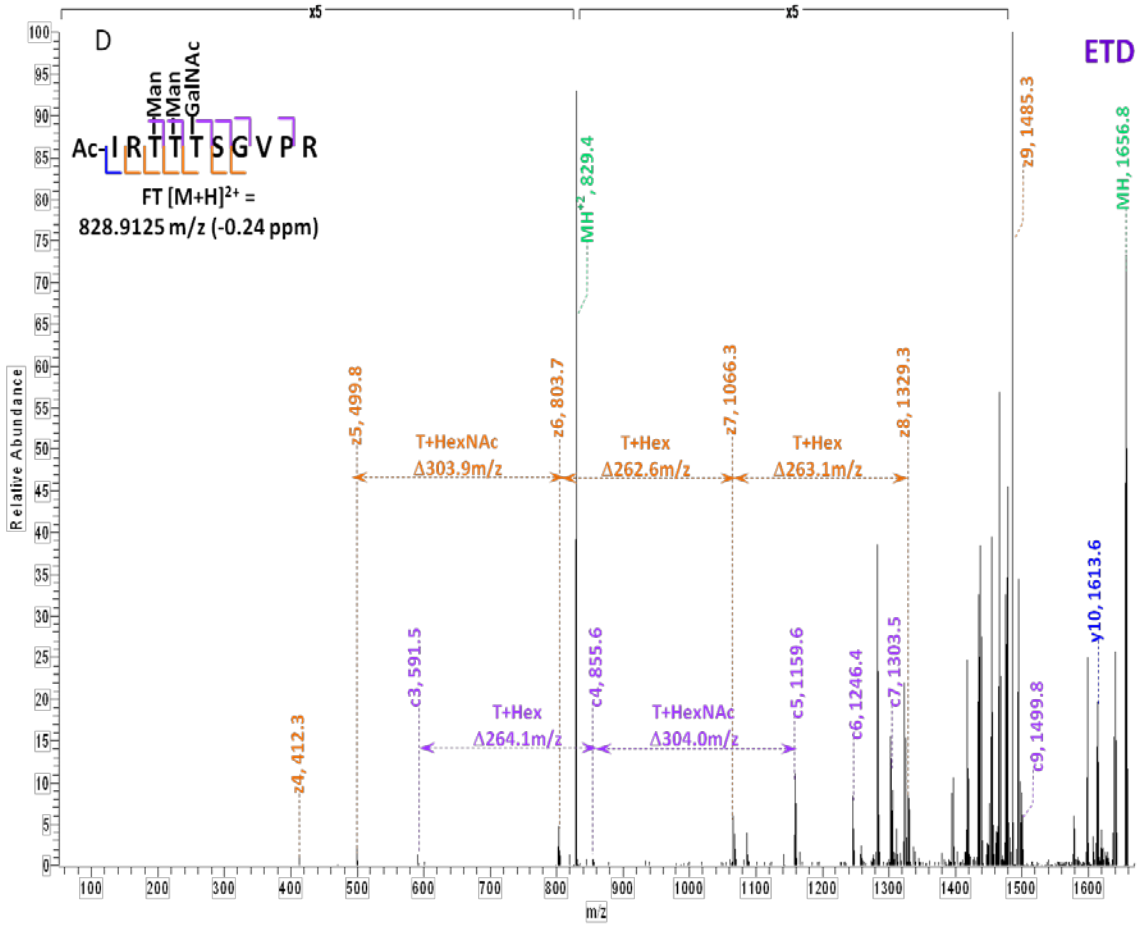


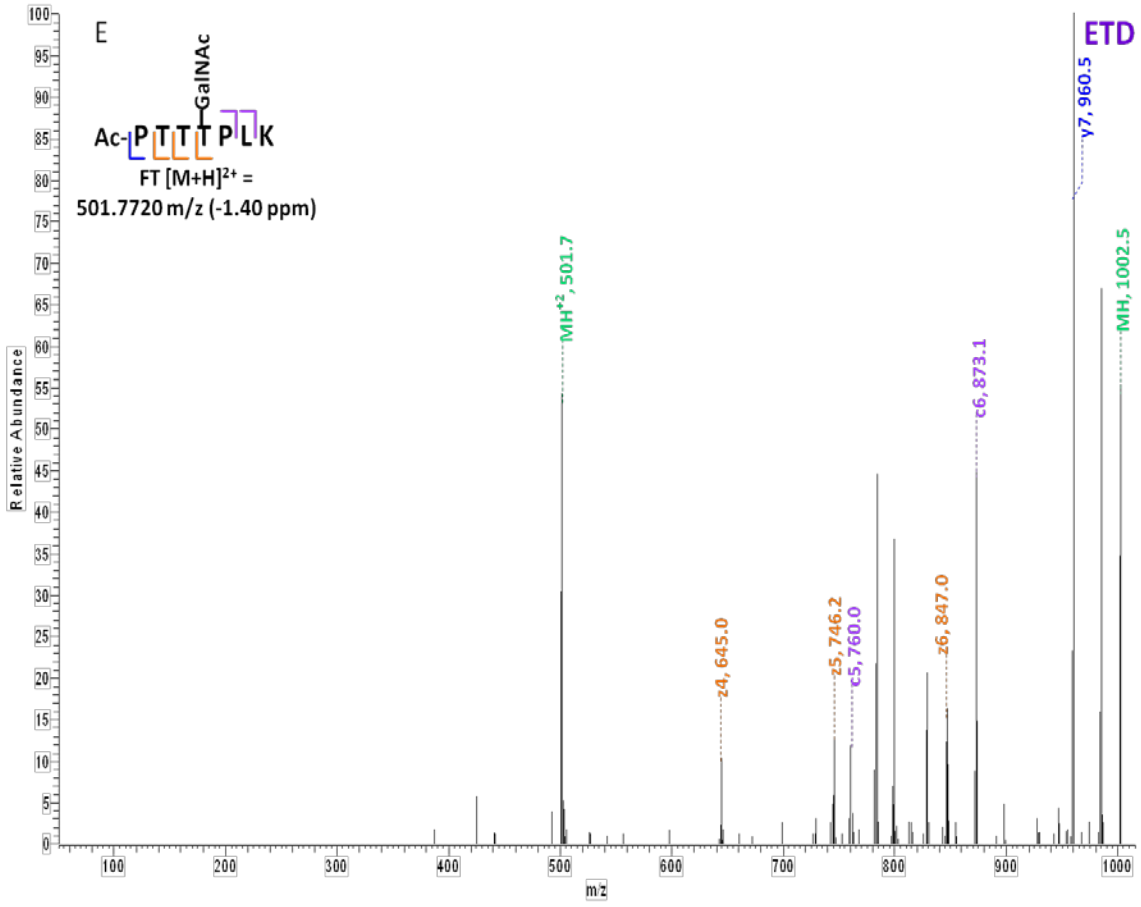
Figure 4-S1. Multiple-engine database search strategy for the HEK293T sample. Raw spectra were searched twice using three search engines: SEQUEST, Mascot and ProteinProspector. The first search was performed against the entire human database without allowing for HexNAc modification, and the output was filtered at 5% peptide-level FDR and exported as the limited protein database; the second search was performed against the exported database allowing for HexNAc modification, and the output was filtered at 1% peptide-level FDR, manually verified and combined to produce a list of HexNAc modified peptides.


A
HCD (zoomed in)
Man
Man
GalNAc
Ac-IRT T TSGVPR
FT [M+H]2+ =
828.9125 m/z (-0.24 ppm)
Relative Abundance
m/z
138.054
[C7H8O2N]+
126.054
[C6H8O2N]+
144.064
[C6H10O3N]+
186.074
[C8H12O4N]+
145.067
[C6H11O5]+
204.085
HexNAc+1
[C8H14O5N]+
115.039
[C5H7O3]+
127.038
[C6H7O3]+
168.064
[C8H10O3N]+
109.027
[C6H5O2]+


B
HCD (zoomed in)
GalNAc
Ac-PTTTPLK
FT [M+H]2+ =
501.7751 m/z (-7.57 ppm)
126.055
[C6H8O2N]+
138.055
[C7H8O2N]+
144.065
[C6H10O3N]+
168.066
[C8H10O3N]+
186.076
[C8H12O4N]+
204.087
HexNAc+1
[C8H14O5N]+
Relative Abundance
m/z


C
HCD (zoomed in)
Man
Ac-RIRTT TSGVPR
FT [M+H]3+ =
483.2715 m/z (-9.31 ppm)
127.040
[C6H7O3]+
109.029
[C6H5O2]+
115.039
[C5H7O3]+
145.050
[C6H11O5]+
163.074
Hex+1
[C6H11O5]+
Relative Abundance
m/z


D
Ac-I R T T T S G V P R
Man
Man
GalNAc
FT [M+H]2+ =
828.9125 m/z (-0.24 ppm)
ETD
x5
x5
Relative Abundance
m/z
MH+2, 829.4
z9, 1485.3
MH, 1656.8
z5, 499.8
z6, 803.7
z7, 1066.3
z8, 1329.3
T+HexNAc
Δ303.9m/z
T+Hex
Δ262.6m/z
T+Hex
Δ263.1m/z
y10, 1613.6
c3, 591.5
c4, 855.6
c5, 1159.6
c6, 1246.4
c7, 1303.5
c9, 1499.8
z4, 412.3
T+Hex
Δ264.1m/z
T+HexNAc
Δ304.0m/z


E
GalNAc
Ac-PTTTPLK
FT [M+H]2+ =
501.7720 m/z (-1.40 ppm)
ETD
y7, 960.5
MH+2, 501.7
MH, 1002.5
c6, 873.1
z4, 645.0
z5, 746.2
c5, 760.0
z6, 847.0
Relative Abundance
m/z

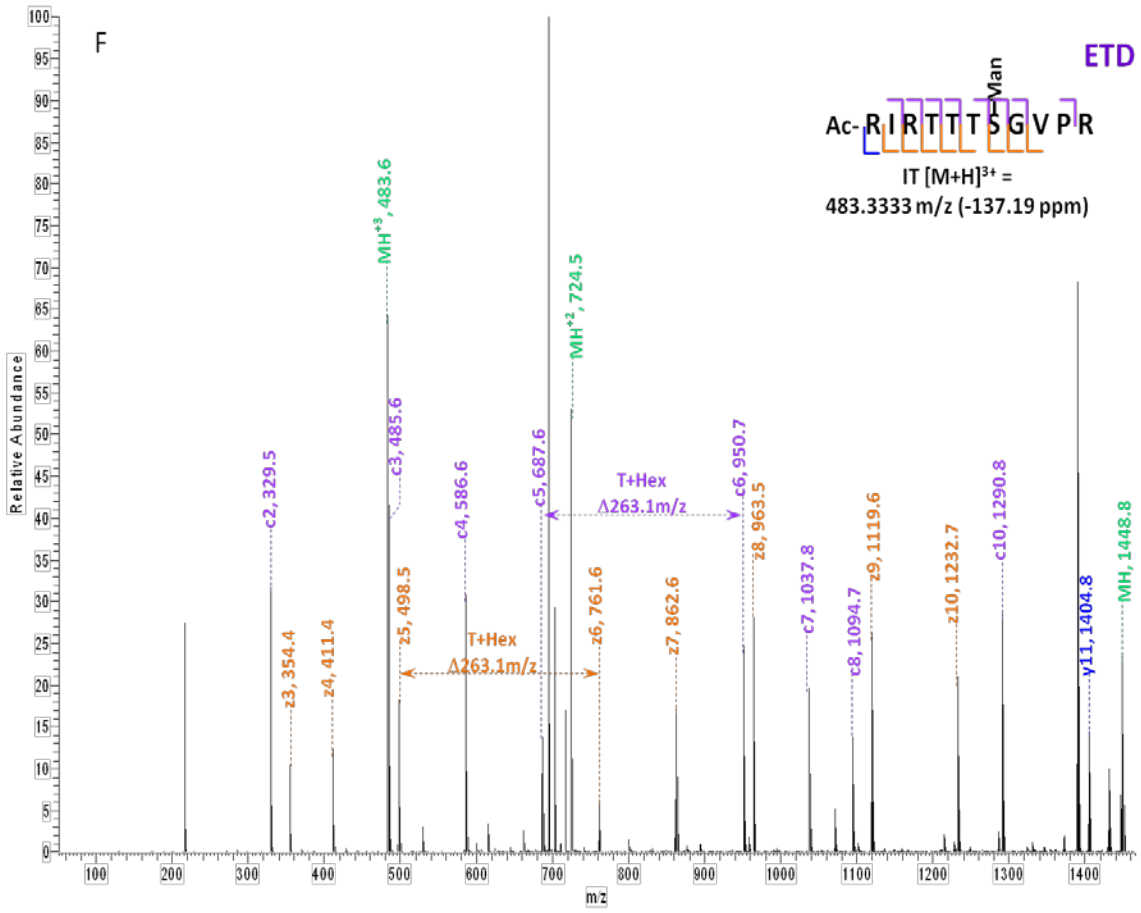


Figure 4-S2. HCD/ETD spectra of O-Mannose and O-GalNAc modified peptides. (**A**) HCD spectrum of peptide Ac-IRt(Man)t(Man)t(GalNAc)SGVPR at low mass range; (**B**) HCD spectrum of peptide Ac-PTTt(GalNAc)PLK at low mass range; (**C**) HCD spectrum of peptide Ac-RIRTTt(Man)SGVPR at low mass range; (**D**) ETD spectrum of peptide Ac-IRt(Man)t(Man)t(GalNAc)SGVPR; (**E**) ETD spectrum of peptide Ac-PTTt(GalNAc)PLK; (**F**) ETD spectrum of peptide Ac-RIRTTt(Man)SGVPR.

Table 4-1. List of novel O-GlcNAc proteins identified in the enriched HEK293T sample that had not been observed in the previous experiment (Ref. 10).

UniProt Accession	Protein Name	Gene Name	Coverage by #AA	# Peptides
B3KVI8	cDNA FLJ16604 fis, clone TESTI4008097, highly similar to Polycomb group protein ASXL1		2.8	5
B4DLF8	cDNA FLJ60295, highly similar to Retinoblastoma-binding protein 5		6.1	5
B4DNK3	cDNA FLJ52127, highly similar to Multisynthetase complex auxiliary component p43		8.55	2
B4DSE4	cDNA FLJ54056, highly similar to Splicing factor 1		8.19	6
B4DUG4	cDNA FLJ51308		23.89	2
B4DW30	cDNA FLJ56743, highly similar to RuvB-like 2 (EC 3.6.1.-)		10.59	3
B4DYD1	Lysozyme C		4.94	2
B4E1W4	cDNA FLJ55383		7.89	4
B4E3U4	cDNA FLJ59107, moderately similar to Heterogeneous nuclear ribonucleoprotein G		10.71	3
B7Z475	cDNA FLJ55712, highly similar to F-box-like/WD repeat protein TBL1XR1		10.64	7
C9J0W3	Putative uncharacterized protein SMARCE1 (Fragment)	SMARCE1	17.2	5
C9JLW8	HCG1818442, isoform CRA_c	hCG_1818442	11.34	6
C9JMM0	Putative uncharacterized protein CBX3	CBX3	48.39	2
C9JU56	Putative uncharacterized protein RPL31	RPL31	35.65	7
O95881	Thioredoxin domain-containing protein 12	TXNDC12	16.86	4
P13929-3	Isoform 3 of Beta-enolase	ENO3	10.49	3
P22061-1	Isoform 1 of Protein-L-isoaspartate(D-aspartate) O-methyltransferase	PCMT1	9.69	10
P24534	Elongation factor 1-beta	EEF1B2	12	5
P25398	40S ribosomal protein S12	RPS12	41.67	7
P26196	Probable ATP-dependent RNA helicase DDX6	DDX6	7.45	8
P30040	Endoplasmic reticulum protein ERp29	ERP29	8.43	2
P30050-1	Isoform 1 of 60S ribosomal protein L12	RPL12	17.58	6
P35268	60S ribosomal protein L22	RPL22	21.88	12
P37108	Signal recognition particle 14 kDa protein	SRP14	16.18	2
P49207	60S ribosomal protein L34	RPL34	22.22	3
P50991	T-complex protein 1 subunit delta	CCT4	5.38	2
P60866	40S ribosomal protein S20	RPS20	19.33	7
P62081	40S ribosomal protein S7	RPS7	21.65	10
P62263	40S ribosomal protein S14	RPS14	18.54	9
P62266	40S ribosomal protein S23	RPS23	15.38	4
P62750	60S ribosomal protein L23a	RPL23A	17.95	8
P62829	60S ribosomal protein L23	RPL23	28.57	7
P62913-2	Isoform 2 of 60S ribosomal protein L11	RPL11	23.73	6
Q07021	Complement component 1 Q subcomponent-binding protein, mitochondrial	C1QBP	15.6	9
Q2L7G6	Heterogeneous nuclear ribonucleoprotein-R2		8.4	4
Q5T4L4	Ribosomal protein S27	RPS27	16.67	3
Q6MZS5	Putative uncharacterized protein DKFZp686A13234 (Fragment)	DKFZp686A13234	8.32	5
Q8IXM2	Uncharacterized potential DNA-binding protein C17orf49	C17orf49	26.16	3
Q8TAQ2-2	Isoform 2 of SWI/SNF complex subunit SMARCC2	SMARCC2	1.77	2
Q92922	SWI/SNF complex subunit SMARCC1	SMARCC1	4.07	5
Q96K80	Zinc finger CCCH domain-containing protein 10	ZC3H10	8.99	3
Q9NPF5	DNA methyltransferase 1-associated protein 1	DMAP1	8.57	4
Q9NYF8-3	Isoform 3 of Bcl-2-associated transcription factor 1	BCLAF1	11.51	6
Q9Y265-1	Isoform 1 of RuvB-like 1	RUVBL1	21.05	16
Q9Y3I0	UPF0027 protein C22orf28	C22orf28	5.94	5

Table 4-2. List of novel HexNAc modification sites identified in the enriched HEK293T sample.

UniProt Accession	Protein Name	Novel Sites
O95487-2	Isoform 2 of Protein transport protein Sec24B	T327; T341
P35658-2	Isoform 2 of Nuclear pore complex protein Nup214	S1202; S1904; S1905; S1907; T1915; S1916
P49790	Nuclear pore complex protein Nup153	T535; S893; S895; S1017; S1018; T1026; T1041
P51610-1	Isoform 1 of Host cell factor	T405; S620; S622; S623; T625; S628; S638; T640; T651; T652; T658
P52594-2	Isoform 2 of Arf-GAP domain and FG repeats-containing protein 1	S291
P52948-4	Isoform 4 of Nuclear pore complex protein Nup98-Nup96	S262; T264
Q14157-4	Isoform 4 of Ubiquitin-associated protein 2-like	S445
Q2KHR3-1	Isoform 1 of Glutamine and serine-rich protein 1	T1271
Q5T6F2	Ubiquitin-associated protein 2	T487; S494
Q5T8P6-3	Isoform 3 of RNA-binding protein 26	S657; S667
Q6MZP7-1	Isoform 1 of Protein lin-54 homolog	T109
Q7Z589	Isoform 1 of Protein EMSY	S228; T264; T272
Q8IWZ3-1	Isoform 1 of Ankyrin repeat and KH domain-containing protein 1	S1817
Q9H4A3-2	Isoform 2 of Serine/threonine-protein kinase WNK1	S1849
Q9Y520	Isoform 7 of BAT2 domain-containing protein 1	S2196

Table 4-S1. Fragments of HexNAc and Hexose oxonium ions.

HexNAc Fragments	Observed m/z	Calculated m/z	Deviation (ppm)	Empirical formula	Details
126	126.05400	126.05496	7.615725712	$[C_6H_8O_2N]^+$	$HexNAc^+$-$2H_2O$-$COCH_2$
138	138.05390	138.05496	7.678101533	$[C_7H_8O_2N]^+$	$HexNAc^+$-$2H_2O$-CH_2O
144	144.06460	144.06552	6.385983266	$[C_6H_{10}O_3N]^+$	$HexNAc^+$-H_2O-$COCH_2$
168	168.06450	168.06552	6.069061637	$[C_8H_{10}O_3N]^+$	$HexNAc^+$-$2H_2O$
186	186.07480	186.07608	6.878906735	$[C_8H_{12}O_4N]^+$	$HexNAc^+$-H_2O
204	204.08530	204.08665	6.614837374	$[C_8H_{14}O_5N]^+$	$HexNAc^+$
Hex Fragments	Observed m/z	Calculated m/z	Deviation (ppm)	Empirical formula	Details
109	109.02900	109.02841	-5.448121474	$[C_6H_5O_2]^+$	Hex^+-$3H_2O$
115	115.03900	115.03897	-0.295552031	$[C_5H_7O_3]^+$	Hex^+-H_2O-CH_2O
127	127.03950	127.03897	-4.203434716	$[C_6H_7O_3]^+$	Hex^+-$2H_2O$
145	145.04970	145.04954	-1.130648222	$[C_6H_9O_4]^+$	Hex^+-H_2O
163	163.07390	163.06010	-84.65590502	$[C_6H_{11}O_5]^+$	Hex^+

CHAPTER 5

CONCLUSION

The overall purpose of this work was to develop and apply mass spectrometry-based methodologies to identify and quantify proteins especially post-translationally modified proteins in complex biological samples.

In Chapter 2, an in-depth proteomic analysis of human embryonic stem cells was performed using multidimensional liquid chromatography in combined with mass spectrometry technology. 3189 proteins were identified by approximately 24,000 peptides, together with the assignment of 492 phosphorylation sites and 68 O-GlcNAc modification sites. Clustering of these identified proteins generated a plethora of information regarding various biological pathways that are operating within stem cells. In particular, analysis of the phosphorylation sites allowed us to infer what kinases are likely to be active within these cells, and inhibition assays were performed on a subset of these kinases to establish the functional roles for the enzymes within the cells. Additionally, 123 proteins were also revealed by investigating the secreted proteome of hES cells, which likely play a role in extracellular matrix formation and remodeling as well as autocrine signaling for self-renewal and maintenance of the undifferentiated state. Finally, by performing in-depth analysis in triplicate, spectral counts with standard deviations were obtained for many of these proteins and post-translationally modified peptides that will allow us to perform relative quantitative analysis between these cells and any derived cell type.

In Chapter 3, an MS-based glycoproteomic approach was adopted to search for biomarker candidates in pancreatic ductal fluids. Our analysis was based on parallel experiments of pancreatic ductal fluid from donor groups sorted on the following criteria: pancreatic cancer, intraductal papillary mucinous neoplasms (IPMN), pancreatitis, and normal pancreas. Our preliminary glycomic data observed increased abundance of N-linked glycans especially the fucosylated structures as well as higher complexity of O-linked glycans in cancer samples. The proteomic study of the fluid samples identified 451 proteins and quantified 47 proteins. The quantitative analysis of protein expression in samples with various diagnoses revealed several potential biomarker candidates, such as the REG family proteins, α-amylase, pancreatic α-amylase, elastase 2A, bile salt-activated lipase, phospholipase A2, pancreatic lipase-related protein 2, and trypsin 1.

In Chapter 4, an MS scheme utilizing HCD and ETD was developed to target the O-GlcNAc modified proteins in biological samples. The diagnostic ion patterns of HexNAc and Hexose revealed under HCD condition can be utilized to selectively target HexNAc and/or Hexose modified proteins, especially when combined with ETD which preserves labile post-translational modification on proteins, the reliability and accuracy in glycoprotein identification and site localization can be greatly improved. The applicability of the combined HCD/ETD MS scheme in characterizing O-GlcNAc modified proteins from a complex biological sample was then investigated and 58 modification sites were subsequently identified. Additionally, with a multiple-engine database search method, we were able to increase the sensitivity of our discovery drastically to reach a total of 165 sites of O-GlcNAc modification. We further proved the capability of the HCD/ETD scheme in characterizing O-GalNAc and/or O-Mannose

modified peptides, and explored its potential in other modified proteins. Along with the advancement of both hardware and software in mass spectrometry, we anticipate that an HCD-trigger-ETD approach will be implemented and realized on a hybrid linear ion trap/Orbitrap platform in the near future.

APPENDIX A

RNA-GUIDED RNA CLEAVAGE BY A CRISPR RNA-CAS PROTEIN COMPLEX[1]

CONTRIBUTION: Used LC-MS/MS tandem mass spectrometry to identify Cas proteins in *P. furiosus*.

[1] Reprinted with permission from Elsevier Inc.
Caryn R. Hale, Peng Zhao, Sara Olson, Michael O. Duff, Brenton R. Graveley, Lance Wells, Rebecca M. Terns, Michael P. Terns
Cell, 139, 945–956, November 25, 2009

ABSTRACT

Compelling evidence indicates that the CRISPR-Cas system protects prokaryotes from viruses and other potential genome invaders. This adaptive prokaryotic immune system arises from the clustered regularly interspaced short palindromic repeats (CRISPRs) found in prokaryotic genomes, which harbor short invader-derived sequences, and the CRISPR-associated (Cas) protein-coding genes. Here, we have identified a CRISPR-Cas effector complex that is comprised of small invader-targeting RNAs from the CRISPR loci (termed prokaryotic silencing (psi)RNAs) and the RAMP module (or Cmr) Cas proteins. The psiRNA-Cmr protein complexes cleave complementary target RNAs at a fixed distance from the 3' end of the integral psiRNAs. In Pyrococcus furiosus, psiRNAs occur in two size forms that share a common 5' sequence tag but have distinct 3' ends that direct cleavage of a given target RNA at two distinct sites. Our results indicate that prokaryotes possess a unique RNA silencing system that functions by homology-dependent cleavage of invader RNAs.

INTRODUCTION

RNAs that arise from the clustered regularly interspaced short palindromic repeats (CRISPRs) found in prokaryotic genomes are hypothesized to guide proteins encoded by CRISPR-associated (cas) genes to silence potential genome invaders in prokaryotes [1]. CRISPRs consist of multiple copies of a short repeat sequence (typically 25 - 40 nucleotides) separated by similarly-sized variable sequences that are derived from invaders such as viruses and conjugative plasmids [1-7]. CRISPR loci are found in nearly all sequenced archaeal genomes and approximately half of bacterial genomes [1, 4, 8]. cas genes are strictly found in the genomes of prokaryotes that possess CRISPRs, frequently in operons in close proximity to the CRISPR loci [1, 8-9]. Over 40 cas genes have been described, a subset of which is found in any given organism [1, 8-9]. The proteins encoded by the cas genes include predicted RNA binding proteins, endo- and exo-nucleases, helicases, and polymerases [1, 8-9]. Recent studies have demonstrated that CRISPRs and cas genes function in invader defense in prokaryotes. Exposure of microorganisms that possess the CRISPR-Cas system to a virus results in the appearance of new virus-derived sequences at the leader-proximal end of CRISPR loci in the genomes of surviving individuals [10-11]. Moreover, the acquisition or loss of invader-specific CRISPR elements or of Cas protein genes has been directly correlated with virus and plasmid resistance or sensitivity, respectively [10-12]. This rapidly evolving immune system influences the ecology of natural microbial populations [7, 13-14].

RNAs from the CRISPR loci are hypothesized to guide the CRISPR-Cas defense response based on their potential to base pair with invading nucleic acids. Available data indicate that entire CRISPR loci are transcribed from the leader region, producing

primary transcripts containing the full set of CRISPR repeats and embedded invader-derived (or guide) sequences [5, 9, 15-18]. These large precursor RNAs are processed into shorter (~60–70 nucleotide) intermediate RNAs that contain individual invader-targeting sequences (~25–40 nucleotides) by Cas endonucleases that cleave within the repeats [12, 19]. However, the ultimate products of the CRISPR loci appear to be smaller RNAs [12, 17-18]. In Pyrococcus furiosus, the most abundant CRISPR RNAs are two species of ~45 nucleotides and ~39 nucleotides [17]. These small, abundant products of the CRISPR loci are thought to be the prokaryotic silencing (psi)RNAs of the CRISPR-Cas RNA silencing pathway [1, 12, 17].

Intriguingly, the protein-mediated functions of the CRISPRCas system are apparently carried out by distinct sets of Cas proteins in different organisms [8]. Six "core" CRISPR-associated genes (cas1 - cas 6) are found in many and diverse organisms, however, most organisms have only a subset of these 6 genes and only cas1 is present in nearly all organisms that appear to possess the system [1, 8]. Furthermore, the core cas genes in a given organism are complemented by one or more sets of additional cas genes: the cse, csy, csn, csd, cst, csh, csa, csm, and cmr genes [8]. These sets are comprised of 2 to 6 CRISPR-associated genes that cosegregate, and are mostly designated for a prototypical organism (e.g., the cse or Cas subtype Escherichia coli genes) [8]. (The cmr (Cas module RAMP) gene set is named for its 4 RAMP (repeat-associated mysterious proteins; see below) gene members.) E. coli K12, for example, has 3 core cas genes and the full set of 5 cse genes (which includes the E. coli subtype member of the core Cas5 gene family, cas5e) [12]. Phylogenetic analyses suggest that the cas genes are distributed by lateral gene

transfer [8-9, 20]. The functional consequences of the differences in the complement of Cas proteins found among organisms are not yet known.

Functional classes have been predicted for many of the Cas proteins based on sequence, but very few of the proteins have been characterized. Only one of the core Cas proteins, Cas6, has a clearly established function which is to process precursor CRISPR RNAs to release individual invader-targeting RNAs [19]. Cas1wasrecentlyshownto be a DNA-specific endonuclease with properties consistent with a role in processing invader DNA into fragments that become incorporated into CRISPR loci [21]. The five E. coli subtype Cas proteins (Cse1-4 and Cas5e [8]) have been shown to form a complex that processes precursor CRISPR RNAs in E. coli (which lacks Cas6) [12]. Many of the Cas proteins are members of the large superfamily of RAMP proteins, which have features of RNA binding proteins [1, 8, 20]. At least a few of the RAMPs (including for example Cas6) have been found to possess previously unpredicted nuclease activity [12, 19, 22]. The Cas proteins are expected to function in various aspects of maintenance of CRISPR gene loci (including addition of new invader-derived elements in response to infection) as well as psiRNA biogenesis and psiRNA-mediated resistance to invaders.

While there is very strong evidence that CRISPR RNAs and Cas proteins function to silence potential invaders in prokaryotes [10-12], the effector complexes and silencing mechanisms of the CRISPR-Cas pathway remain unknown. Recent studies in Staphylococcus species and E. coli [12, 23] indicate that the CRISPR-Cas systems present in those organisms (comprised of the Csm or Cse proteins and several core Cas proteins, respectively) target invader DNA rather than RNA, but the effectors and mechanisms of silencing in these organisms remain unknown. The results presented here demonstrate

that the Cmr or RAMP module proteins function with mature psiRNAs to cleave target RNAs. These findings define psiRNA-guided RNA cleavage as a mechanism for the function of the CRISPR-Cas system in organisms that possess the RAMP module of Cas proteins.

RESULTS

Isolation of a Complex Containing Mature psiRNAs and a Subset of Cas Proteins

PsiRNAs are hypothesized to guide Cas proteins to effect invader silencing in prokaryotes [1, 12, 17]. P. furiosus is a hyperthermophilic archaeon whose genome encodes 200 potential psiRNAs (organized in seven CRISPR loci) and at least 29 potential Cas proteins (largely found in 2 gene clusters), including members of all six core Cas protein families and three sets of additional Cas proteins: the Cmr, Cst and Csa proteins (see Figure A-1F). In P. furiosus, most psiRNAs are processed into two species of ~45 nucleotides and ~39 nucleotides [17]. To gain insight into the functional components of the CRISPR-Cas invader defense pathway, we isolated complexes containing the mature psiRNA species from P. furiosus cellular extract on the basis of psiRNA fractionation profiles (Figure A-1). The doublet of psiRNAs, detectable both by Northern blotting of an individual psiRNA and total RNA staining (SYBR), was purified away from larger CRISPR-derived RNAs (including the 13 intermediate) [17] as well as other cellular RNAs (Figure A-1C).

To determine whether the psiRNAs are components of RNA-protein complexes in the purified fraction (Figure A-1C), we performed native gel northern analysis. The mobility of the psiRNAs on native gel electrophoresis was reduced in the purified fraction relative to a sample from which proteins were extracted (Figure A-1D),

indicating the presence of psiRNA-protein complexes in the purified fraction. We gel purified the psiRNA-containing complex from the native gel and analyzed the sample by mass spectrometry. The sample contained a mixture of proteins that included seven Cas proteins identified with 99% confidence: Cmr1-1, Cmr1-2, Cmr2, Cmr3, Cmr4, Cmr5, and Cmr6 (Figure A-1E).

The identities of the non-Cas proteins found in the sample are listed in Table A-S1, available online. Analysis of a native gel-purified psiRNP obtained by analternate chromatography scheme revealed a similar Cas protein profile (Cmr2,Cmr3,Cmr4, andCmr6), but few common non-Cas proteins (Table A-S1). The five common copurifying non-Cas proteins are denoted in Table A-S1. None of these proteins has any known link to the CRISPR-Cas system.

Remarkably, the seven Cas proteins associated with the complex are all encoded by the tightly linked RAMP module or cmr genes [8]. Moreover, the identified proteins comprise the complete set of Cmr proteins [8]. (The independently defined "polymerase cassette" is closely related to the RAMP module [1].) There are 6 cmr genes: cmr2 encodes a predicted polymerase with HD nuclease domains, and cmr1, cmr3, cmr4, and cmr6 encode repeat-associated mysterious proteins (RAMPs) [8, 20]. The P. furiosus genome contains two cmr1 genes and a single representative of each cmr2 – cmr6, and all seven corresponding proteins were found in the purified psiRNP complex (Figure A-1E). The organization of the genes encoding the seven identified proteins is shown in Figure A-1F. Six of the seven identified Cas proteins are encoded in a nearly contiguous region of one of the two major cas gene loci in P. furiosus. This locus is located directly adjacent to

CRISPR locus 7, and also encodes core Cas proteins Cas1 - Cas4, Cas5t, and Cas6. The striking correlation between the evolutionary cosegregation and physical association of the 6 Cmr proteins strongly supports the cofunction of the proteins. Our findings indicate that the two mature psiRNA species are components of complexes containing the RAMP module or Cmr proteins in P. furiosus.

psiRNAs Possess a 5' psiRNA-Tag Sequence

In order to better understand the nature of the two psiRNA species that are components of the purified complexes, each of the two RNA bands present in the final chromatography sample (Figure A-2A) was extracted and cloned. We obtained sequences of 51 RNAs (20 from the upper band and 31 from the lower band) that included psiRNAs from all seven P. furiosus CRISPR loci. Six RNAs with the same guide sequence were represented in both the upper and lower bands, consistent with Northern analysis that has shown that most psiRNAs exist in both size forms [17].

The cloned psiRNAs consisted primarily of an individual guide (invader-targeting or "spacer") sequence, however, all of the clones retained a portion of the common repeat sequence at the 5' end. Indeed, the majority (~70%) of the RNAs in both bands contained an identical 5' end consisting of an 8-nucleotide segment of the repeat sequence (Figure A-2A). The difference between the two psiRNA size forms was found at the 3' ends. Downstream of the repeat sequence, the majority of the clones from the top band contained 37 nucleotides of guide sequence (the full length of a typical guide element in P. furiosus) (Figure A-2A, top panel). The 3' ends of most of the clones from the bottom band were located within the guide sequence. The majority of these RNAs contained 31

nucleotides of guide sequence downstream of the repeat sequence (Figure A-2A, bottom panel).

The psiRNAs are processed from long CRISPR locus transcripts [5, 12, 15-18] (Figure A-2B). In P. furiosus, the Cas6 endoribonuclease cleaves CRISPR RNAs at a site within the repeat element located 8 nucleotides upstream of the guide sequence, generating the precise 5' end observed in the two psiRNA species found in the complex (Figure A-2B) [19]. Our results indicate that the 5' end generated by the Cas6 endoribonuclease is maintained in the mature psiRNAs, but that the RNAs undergo further processing at the 3' end to generate psiRNAs that contain either ~37 or ~31 nucleotides of guide sequence (Figure A-2B). Themechanismthat defines the two distinct 3' end boundaries is not known. The larger ~45-nucleotide mature psiRNA species is generally more abundant than the smaller ~39-nucleotide species [17] (Figures A-1 and A-2A).

The short repeat sequence that remains at the 5' end of mature psiRNAs in P. furiosus provides a common identifying sequence tag for the psiRNAs that could function in recognition of the RNAs by the proteins in the CRISPR-Cas pathway. In order to more rigorously delineate the potentially important psiRNA-tag or "psi-tag," we purified small RNAs from P. furiosus, performed deep sequencing and obtained the sequences of the 5' ends of more than 10,000 CRISPR-derived RNAs (from loci 1-7). The 5' ends of the majority of the RNAs mapped 8 nucleotides upstream of the guide sequence (Figure A-2C), verifying the presence of a discrete psi-tag on small CRISPR-derived RNAs in P. furiosus.

The sequences of CRISPR repeats (from which psi-tags are derived) are generally conserved within groups of organisms, but can vary widely [4, 24]. Thus, while the sequence

of the psi-tag found on most P. furiosus psiRNAs (AUUGAAAG) can be found in the repeat sequence of numerous organisms, psi-tags of distinct sequence and length would be expected in others. We found evidence to support this prediction in the psiRNAs from P. furiosus CRISPR locus 8, which contains a single nucleotide deletion in the psi-tag region of the repeat. The majority (60%) of the 640 sequenced RNAs that mapped to CRISPR locus 8 possessed a 7-nucleotide AUUGAAG psi-tag. In E. coli, CRISPR transcripts are cleaved by a different endoribonuclease (Cse3 of the Cse complex), which nonetheless appears to generate RNAs with an 8-nucleotide AUAAACCG repeat sequence at the 5' end [12]. An 8-nucleotide ACGAGAAC repeat sequence is also present at the 5' termini of CRISPR RNAs in S. epidermidis [23], suggesting that the psi-tag is a general feature of the psiRNAs. Interestingly, the distinct CRISPR repeat sequences found in various genomes are accompanied by distinct subsets of Cas proteins [24], which may reflect coupling of specific series of Cas proteins with the psi-tagged RNAs that they recognize.

Homology-Dependent Cleavage of a Target RNA

One hypothesis for the mechanism by which CRISPR RNAs and Cas proteins mediate genome defense is psiRNA-guided cleavage of invader nucleic acids [1]. Therefore, we tested the ability of the isolated psiRNP complexes to recognize and cleave a labeled RNA and DNAtarget complementary to endogenous P. furiosus psiRNA 7.01 (first psiRNA encoded in CRISPR locus 7, which Northern analysis indicated is present in the native complexes, see Figure A-1). The 5' end-labeled 7.01 target RNA was cleaved at two sites (site 1 indicated with green vertical line and site 2 indicated with blue vertical line, substrate 1, Figure A-3B) yielding 5' end-labeled products of 27 and 21

nucleotides (indicated with corresponding green and blue arrowheads, substrate 1, Figure A-3A). The single-stranded DNA 7.01 target sequence was not cleaved (substrate 3, Figure A-3).

Further characterization of the cleavage activity revealed that the psiRNP complexes cleave the target RNA on the 5' side of the phosphodiester bond. The 3' end generated by the complex is not a substrate for polyadenylation, indicating the presence of a 3' phosphate (or 2', 3' cyclic phosphate) end. In addition, cleavage activity is lost in the presence of 0.1 mM EDTA indicating that the enzyme depends on divalent cations. Activity was restored by the addition of 1 mM Mg2+, Mn2+, Ca2+, Zn2+, Ni2+ or Fe2+ with no detectable change in cleavage sites with any of the metals, but was not supported by Co2+or Cu2+. Cleavage of the target RNA did not require sequences extending beyond the 37-nucleotide region of complementarity with the psiRNA, and occurred at the same two sites in the target RNA lacking sequence extensions (substrate 6, Figure A-3). No activity was observed toward RNAs that lacked homology with known P. furiosus psiRNAs, including the reverse 7.01 target sequence, antisense 7.01 target sequence, and a box C/D RNA (substrates 2, 7 and 8, Figure A-3). Pre-annealing a synthetic psiRNA 7.01 to the 7.01 target RNA (to form a double-stranded RNA target) blocked cleavage by the psiRNPs (substrate 5, Figure A-3). Finally, we tested a target for endogenous P. furiosus psiRNA 6.01 and observed cleavage that generates 2 products of the same sizes observed for the 7.01 target RNA (substrate 4, Figure A-3).

These results demonstrate the presence of cleavage activity in P. furiosus that is specific for single-stranded RNAs that are complementary to psiRNAs. The activity is

associated with a purified fraction that contains 2 mature psiRNA species and 7 RAMP module (Cmr) proteins.

Cleavage of the Target RNA Occurs a Fixed Distance from the 3' End of the psiRNA

To investigate the mechanism of psiRNA-directed RNA cleavage, we analyzed the results of cleavage assays with a series of truncations of the 7.01 target RNA (Figure A-4A). We found that the target RNA truncations analyzed did not affect the locations of the two cleavage sites. The full-length 7.01 target RNA is cleaved at sites 1 and 2 to generate 14- and 20-nucleotide 5' end-labeled products, respectively (Figures A-3 and A-4A). The 3' end-truncated target RNAs were cleaved at the same two sites to yield the same two 5' end-labeled cleavage products (except where truncation eliminated cleavage site 2, Δ20–37, Figure A-4A). On the other hand, in the case of the 5' end-truncated target RNAs, cleavage at the same sites would be expected to generate shorter 5' end-labeled cleavage products. The 14-nucleotide product that results from cleavage of the Δ1-6 target RNA at site 2 was observed (Figure A-4A), but cleavage at site 1 could not be assessed because the size of the product is below that which could be detected in the experiment. If the twelve and eighteen-nucleotide 5' end-truncated target RNAs were cleaved at the same two sites, the products would also be outside the range of detection, however, interestingly, very little cleavage of these RNAs was observed (Figure A-4, Δ1–18 and Δ1–12, compare substrate band +/- complex).

Strikingly, the difference in the sizes of the two cleavage products observed with the various substrates is the same as the difference in the sizes of the two endogenous psiRNA species (6 nucleotides in both cases, Figure A-3). This size difference as well as the specific product sizes suggest that the two cleavages occur a fixed distance (14

nucleotides) from the 3’ ends of the two psiRNAs. Figure A-4B illustrates the proposed mechanism by which the 45- and 39-nucleotide psiRNAs guide cleavage at target sites 1 and 2, respectively, for each of the target RNAs analyzed here. For example, using the full-length 7.01 target RNA we observed 20- and 14-nucleotide cleavage products (Figure A-3, panel 5) suggesting cleavage of the bound target RNA 14 nucleotides from the 3’ end of the 39- and 45-nucleotide psiRNAs, respectively (Figure A-4B, F.L.). In addition, a 7-nucleotide extension at the 5’ end of the target RNA resulted in a pair of 5’ end-labeled products 27 and 21 nucleotides in length (Figure A-3A), consistent with cleavage of the substrate 14 nucleotides from the ends of the two psiRNAs (Figure A-4B, F.L.+ext). The anchor for this counting mechanism is the 3’ end of the psiRNA. While reductions in the extent of duplex formation between the 5’ end of the psiRNA and the cleavage site (3’ truncations to within 6 nucleotides of the cleavage site) did not have an observable effect on cleavage efficiency, truncations that reduced duplex formation between the 3’ end of the psiRNA and the cleavage site had a strong negative impact, suggesting that base-pairing of the last 14 nucleotides of the psiRNA with the target is critical for cleavage activity.

The results of these studies indicate that both of the mature psiRNA species are active in guiding target RNA cleavage by a mechanism that depends upon the distance from the 3’ end of the psiRNA.

Analysis of Reconstituted Cmr-psiRNA Complexes

Identification of the Cmr proteins in the purified psiRNP complex (Figure A-1) along with the evolutionary evidence for their cofunction with the CRISPRs [8-9, 20] strongly suggests that the Cmr proteins and psiRNAs function as a complex to cleave

target RNAs (Figure A-3). In order to determine whether the Cmr proteins and psiRNAs are sufficient for function (independent of other copurifying P. furiosus components), we tested the ability of purified recombinant Cmr proteins and synthetic psiRNAs to cleave target RNAs (Figure A-5). A reconstituted set of six P. furiosus Cmr proteins (Cmr1-1, Cmr2 – Cmr6) and two mature psiRNA species (45- and 39-nucleotide psiRNA 7.01, found in the native complex based on Northern analysis [Figure A-1] and activity of the native complex against the 7.01 target [Figure A-3]) cleaved the target RNA at 2 sites generating the same size products as those observed with the isolated native complex (Figure A-5A). While both P. furiosus isoforms of the Cmr1 protein are present in the isolated complexes (Figure A-1), we found that only one of the two proteins (Cmr1-1) was required for a functional reconstituted complex (Figure A-5A), suggesting that the isoforms may perform redundant functions. No activity was observed in the absence of the psiRNAs or in the absence of the Cmr proteins (Figure A-5A), indicating that both are necessary. These results demonstrate that the RAMP module Cas proteins and psiRNAs function together to cleave complementary target RNAs.

In order to determine whether all of the six Cmr proteins are essential for psiRNA-guided RNA cleavage, we assayed cleavage activity in the absence of each of the individual proteins (Figure A-5B). Omission of Cmr5 did not observably affect the activity of the complex (Figure A-5B). However, cleavage was significantly reduced in the absence of any one of the other proteins (Figure A-5B), indicating that 5 of the 6 RAMP module proteins are required for activity of the psiRNA-Cmr protein complex.

Finally, we had reconstituted the cleavage activity profile observed for the native complexes using the two psiRNA species (45- and 39-nucleotides) (e.g., Figure A-5A).

Our model for the mechanism of cleavage predicts that each of the psiRNAs guides a distinct cleavage: the 45-nucleotide psiRNA at site 1, and the 39-nucleotide psiRNA at site 2 (see Figure A-4B). To determine whether both psiRNAs are required for activity, and whether each guides the distinct cleavage that is predicted by the model, we tested the activity of complexes reconstituted with a single psiRNA. As predicted, we found that the 45-nucleotide psiRNA guided cleavage at site 1 producing a 14-nucleotide 5' end-labeled product, and the 39-nucleotide psiRNA guided cleavage at site 2 producing a 20-nucleotide 5' end-labeled product (Figure A-5C). Based on our truncation analysis (Figure A-4, Δ20-37), the larger product of the cleavage guided by the 39-nucleotide psiRNA could act as a substrate for cleavage guided by the 45-nucleotide psiRNA, and consistent with this, we often obtain more of the smaller cleavage product in cleavage assays where both guide RNAs are present with either the native complex or the reconstituted complex (e.g., Figure A-5A). The results of these experiments demonstrate that each of the psiRNA species is competent to form functional psiRNPs and guides cleavage 14 nucleotides from its 3' end.

DISCUSSION

The findings presented here reveal the mechanism of action of an RNA-protein complex implicated in a novel RNA silencing pathway that functions in invader defense in prokaryotes. Previous work had shown that both invader-specific sequences within CRISPRs and Cas protein genes are important in virus and plasmid resistance in prokaryotes [10-12, 23]. The results presented here establish how small RNAs from CRISPRs and the RAMP module Cas proteins function together to destroy RNAs recognized by the

CRISPR RNAs. The major findings and models established in this work are summarized in Figure A-6.

Our findings indicate that the RAMP module of the CRISPRCas system silences invaders by psiRNA-guided cleavage of invader RNAs (Figure A-6). Specifically, the results indicate that psiRNAs present in complexes with the Cmr proteins recognize and bind an invader RNA such as a viral mRNA (via the psiRNA guide sequence co-opted from the invader by another branch of the CRISPR-Cas system), and that the complex then cleaves the invader RNA, destroying the message and presumably blocking the viral life cycle. The psiRNA-Cmr complexes cleave complementary RNAs (Figures 3 and 5). Five of the six Cmr proteins are required for target RNA cleavage (Figure A-5) and the component of the complex that provides catalytic activity remains to be determined. Cmr2 contains a predicted nuclease domain [1, 20], however the other four essential proteins (Cmr1, 3, 4 and 6) belong to the RAMP superfamily, members of which have been found to be ribonucleases [12, 19, 22]. It will be important in future work to identify the catalytic component(s) of the psiRNA-Cmr protein complex. Our data indicate that the Cmr ribonuclease generates products with 3' phosphate (or 2', 3' cyclic phosphate) and 5' hydroxy termini and requires divalent metal ions for activity.

Our results also establish a simple model for the mechanism of cleavage site selection by the psiRNA-Cmr effector complex - a 14-nucleotide ruler anchored by the 3' end of the psiRNA (Figure A-6). We found that P. furiosus psiRNAs occur in two lengths that share a 5' psi-tag (derived from the CRISPR repeat) and contain either ~37 or ~31 nucleotides of guide sequence (Figures 1 and 2). Both psiRNA species are associated with the Cmr effector complex (Figure A-1) and each guides cleavage at a distinct site

(Figure A-5C). Analysis of the cleavage products of both psiRNAs and of a series of substrate RNAs (Figures 3, 4, and 5) indicates that the complex cleaves based on a 14-nucleotide counting mechanism anchored by the 3' end of the psiRNA. The results suggest that the 3' end of the psiRNA places the bound target RNA relative to the enzyme active site (Figure A-6).

The activity of the psiRNA-Cmr protein complex (RNA-guided RNA cleavage) bears an interesting resemblance to that of Argonaute 2 (a.k.a. Slicer) [25], an enzyme with an analogous function in the eukaryotic RNAi pathway, however there is little similarity between the enzymes. There is no significant sequence homology between the Cmr proteins and Argonaute 2 (or between any of the Cas proteins and known components of the eukaryotic RNAi pathway). Both the psiRNA-Cmr complex and Argonaute 2 employ a ruler mechanism for cleavage site selection; however, in the case of Argonaute 2, the site of cleavage is located ~10-11 nucleotides from the 5' end of the siRNA [26-27]. The activity of both enzymes requires divalent metal ions [28], however for the psiRNA-Cmr RNP, it is not yet clear whether the metal is involved in cleavage catalysis or is required for some other essential aspect of the functionality of this multi-component complex. Finally, Argonaute 2 cleaves target RNAs on the 3' side of the phosphodiester bond, leaving 3' OH and 5' phosphate termini [29]. It is interesting that eukaryotes and prokaryotes exploit distinct small RNA-guided gene silencing pathways to combat viruses and other mobile genetic elements that they encounter [30-31].

Figure A-6 also illustrates the Cmr-psiRNA effector complex model that arises from the findings presented here. Both size classes of psiRNAs and all seven Cmr proteins are found in complexes in active, purified fractions (Figure A-1), however

accurate RNA-guided cleavage activity can be reconstituted with either psiRNA species and with a single Cmr1 isoform (Figure A-5). We hypothesize that each psiRNA associates with a single set of six Cmr proteins, and that Cmr1-1 and Cmr1-2 function redundantly in P. furiosus. Five unrelated proteins that copurified with the complexes (Table A-S1) are not essential for reconstitution of cleavage activity in vitro (Figure A-5) and are not included in our model, but could play a role in function in vivo. Recognition of the psiRNAs by the Cmr proteins and psiRNA-Cmr complex assembly likely depend upon conserved features of the RNAs that could include 5' and 3' end groups and folded structure as well as the psi-tag. Our data reveal that the psiRNA-Cmr complex can utilize psiRNAs of different sizes to cleave a target RNA at distinct sites (Figure A-5C). Thus, the two size forms of psiRNAs present in P. furiosus may provide more certain and efficient target destruction.

Our data indicate that the function of the RAMP module of Cas-proteins is psiRNA-guided destruction of invading target RNA. The widespread occurrence of the cmr genes in diverse archaea (including Sulfolobus and Archaeoglobus species) and bacteria (including Bacillus and Myxococcus species) indicates that invader RNA cleavage is a mechanism utilized by many prokaryotes for viral defense [1, 8-9]. However, not all prokaryotes with the CRISPR-Cas system possess the RAMP module (Cmr) proteins. In these numerous other organisms, it is possible that a different set of Cas proteins mediates psiRNA-guided RNA cleavage or that Cas proteins effect invader resistance by another mechanism. Indeed, very recent work indicates that the CRISPR-Cas system targets invader DNA in a strain of Staphylococcus epidermidis and perhaps E. coli [12, 23], which possess the Mtube (Csm) and Ecoli (Cse) subtype Cas protein modules,

respectively [1, 8-9]. The prokaryotes include evolutionarily distant and very diverse organisms. Diversity in the core components of the eukaryotic RNAi machinery has led to a tremendous variety of observed RNA-mediated gene silencing pathways that can act at post-transcriptional or transcriptional levels [32-35]. The diversity of Cas proteins found in CRISPR-containing prokaryotes may reflect significantly different mechanisms of CRISPR element integration, CRISPR RNA biogenesis, and invader silencing.

EXPERIMENTAL PROCEDURES

Chromatography

P. furiosus S100 extract was prepared from approximately 4 g of cells. Cells were resuspended in 20 ml of 50 mM Tris (pH 7.0), 100 U RNase-free DNase (Promega), and 0.5 mM phenylmethanesulphonyl fluoride (PMSF) at room temperature by stirring. The resulting whole-cell extract wassubject to ultracentrifugation at 100,000 x g for 1.5 hr using an SW 41 Ti rotor (Beckman). The resulting S100 extract was loaded onto a 5 ml Q-sepharose Fast Flow (GE) pre-packed column. Proteins were eluted using a 0-1 M NaCl gradient. Fractions were analyzed by Northern analysis by isolating RNA from 100 ul of each fraction using Trizol LS (Invitrogen, following manufacturer's instructions). The RNAs were separated on 15% TBE-urea gels (Criterion, Bio-Rad), blotted and analyzed for the presence of a single guide sequence as described previously [17]. Peak fractions containing the psiRNA doublet were further separated on a second 5 ml Q-sepharose column, eluted with 220-430 mM NaCl. Fractions were analyzed as described above. Peak fractions were pooled, diluted in 50 mM sodium phosphate buffer, pH 7.0, and loaded onto a 5 ml S-sepharose column (GE). Bound proteins were eluted with a gradient of 0-1M NaCl. Native gel northern analysis was performed as described

previously [17]. The secondary data shown in Table A-S1 was obtained from S100 extract fractionated on a DEAE column as previously described [36] followed by a hydroxyapatite column eluted with a gradient of 5–500 mM sodium phosphate buffer (pH 6.5) and further purified by native gel electrophoresis.

Protein Assignment by Tandem Mass Spectrometry

In-gel and in-solution tryptic digests were performed as previously described [37-38]. Desalted tryptic peptides were analyzed by nLC-MS/MS on a linear ion-trap (LTQ, ThermoFisher) as previously described [38]. Acquired data was searched against a P. furiosus-specific database (forward and inverted) using the TurboSEQUEST algorithm (Thermo-Fisher). Data was collated and filtered to obtain a1%false discovery rate at the protein level using the ProteoIQ software package (BioInquire) that is based on the PROVALT algorithm [39].

Cloning and Sequencing of psiRNAs from the Purified Complexes

RNAs from S-column fractions (isolated as described above for Northern analysis) were treated with 1 U calf intestinal alkaline phosphatase (Promega) for 1 hr at 37°C, followed by extraction with phenol:chloroform:isoamyl alcohol (PCI; [pH 5.2], Fisher) and ethanol precipitation. The resulting RNAs were separated by 15% polyacrylamide, TBE-urea gels (Criterion, Bio-Rad), visualized by SYBR Gold staining (Invitrogen) and the visible bands were excised. RNAs were passively eluted overnight in 0.5 M ammonium acetate, 0.1% SDS, 0.5 mM EDTA, followed by ethanol precipitation. A 5'-phosphorylated, 3' capped oligonucleotide (5'-pCTCGAGATCTGGATCCGGG-ddC3'; IDT) was ligated with T4 RNA ligase to the 3' end of the RNAs. The ligated RNAs were PCI extracted, ethanol precipitated, gel purified, and subject to reverse

transcription using Superscript III (Invitrogen) RT (as described by the manufacturer), followed by gel purification. The gel-purified cDNAs were polyA-tailed for 15 min at 37 °C using terminal deoxynucleotide transferase (Roche) using manufacturer's recommendations. PCR was performed to amplify the cDNA libraries using the following primers: 5'-CCCGGATCCAGATCTCGAG-3', 5'-GCGAATTCTGCAG$(T)_{30}$-3'). cDNAs were cloned into the TOPO pCRII (Invitrogen) cloning vector and transformed into TOP10 cells. White and light-blue colonies were chosen for plasmid DNA preparation, and sequencing using the M13 Reverse and T7 promoter sequencing primers was performed by the University of Georgia Sequencing and Synthesis Facility.

Small RNA Deep Sequencing

Small RNA libraries were prepared using the Illumina small RNA Sample preparation kit as described by the manufacturer (Illumina). Briefly, total RNA was isolated from P. furiosus and fractionated on a 15% polyacrylamide/urea gel, and small RNAs 18-65 nt in length were excised from the gel. 5' and 3' adapters were sequentially ligated to the small RNAs and the ligation products were gel-purified between each step. The RNAs were then reverse-transcribed and PCR-amplified for 16 cycles. The library was purified with a QIAGEN QuickPrep column and quantitated using an Agilent Bioanalyzer and a nanodrop. The sample was diluted to a concentration of 2 pM and subjected to 42 cycles of sequencing on the Illumina Genome Analyzer II.

Small RNA Analysis

Sequence data was extracted from the images generated by the Illumina Genome Analyzer II using the software applications Firecrest and Bustard. The adaptor sequences were then trimmed from the small RNA reads, which were then mapped to the P. furiosus

genome using btbatchblast. Only reads that mapped perfectly to the genome over their entire length were used for further analysis. The location and number of reads that initiate within the CRISPR repeats were determined using a perlscript. As the maximal read length of the sequences was 42 nt, it was not possible to be certain that the 3' end of a read represented the actual 3' end of the small RNA. Therefore, the deep sequencing data was only used to determine the 5' ends.

Nuclease Assays

To detect target RNA cleavage, 2 ml of the peak S-column fractions (Figure A-1C) or 500 nM each of recombinant proteins was incubated with 0.05 pmoles of ^{32}P-5' end-labeled synthetic target RNAs (Figures 3, 4, and 5) and 0.5 pmoles of each unlabeled psiRNA (Figure A-5) for 1 hr at 70 °C in 20 mM HEPES (pH 7.0), 250 mM KCl, 1.5 mM MgCl2, 1 mM ATP, 10 mM DTT, in the presence of 1 unit of SUPERase-In ribonuclease inhibitor (Applied Biosystems). For assays with recombinant proteins, the psiRNAs were first incubated with the proteins for 30 min at 70 °C prior to the addition of target RNA. Reaction products were isolated by treatment with 800 ng of proteinase K for 30 min at room temperature, followed by PCI extraction and ethanol precipitation. The resulting RNAs were separated on 15% polyacrylamide, TBE 7M urea gels and visualized by phosphorimaging. 5' end-labeled RNA size standards (Decade Markers, Applied Biosystems) were used to determine the sizes of the observed products. Annealed RNAs were prepared by mixing equimolar amounts of RNAs in 30 mM HEPES (pH 7.4), 100 mM potassium acetate, 2 mM magnesium acetate and incubating for 1 min at 95 °C, followed by 1 hr at 37 °C. Annealing was confirmed by non-denaturing 8% PAGE.

For analysis of the chemical ends of the cleavage products, cleavage reactions were performed using 5'-end labeled target as described above. The resulting RNA products were isolated by PCI extraction and ethanol precipitation, and subject to polyadenylation by incubation with 5 U E. coli polyA polymerase (NEB) for 15 min at 37 °C as described by the manufacturer. The reaction was stopped by PCI extraction, followed by ethanol precipitation. The resulting products were analyzed on 15% polyacrylamide, TBE 7M Urea gels as described above.

In order to determine the divalent metal requirements of the purified complex, cleavage reactions were performed for 1 hr at 70 °C in 50 mM HEPES (pH 7.0), 250 mM KCl, 1 mM ATP, 10 mM DTT, 0.1 mM EDTA, and 1 mM metal (if applicable) in the presence of 1 unit of SUPERase-In ribonuclease inhibitor (Applied Biosystems). Certified metal reference solutions (Spex CertiPrep except calcium obtained from Fisher Scientific) were added to 1 mM final concentration. The resulting products were isolated and analyzed as described above.

Expression and Purification of Recombinant Proteins

The genes encoding P. furiosus Cmr1-1 (PF1130), Cmr2 (PF1129), Cmr3 (PF1128), Cmr4 (PF1126), Cmr5 (PF1125) and Cmr6 (PF1124) were amplified by PCR from genomic DNA or existing constructs and cloned into a modified version of pET24d (PF1124, PF1125, and PF1126) or pET200D (PF1128, PF1129, and PF1130). The recombinant proteins were expressed in E. coli BL21-RIPL cells (DE3, Stratagene). The cells (400 ml cultures) were grown to a OD_{600} of 0.7, and expression of the proteins was induced with 1 mM isopropyl-b-D-thiogalactopyranoside (IPTG) overnight at room temperature. The cells were pelleted, resuspended in 20 mM sodium phosphate buffer

(pH 7.6), 500 mM NaCl and 0.1 mM phenylmethylsulfonyl fluoride (PMSF), and disrupted by sonication. The sonicated sample was centrifuged at 4,500 rpm for 15 min at 4°C. The supernatant was heated at 75°–78°C for 20 min, centrifuged at 4500 rpm for 20 min at 4°C, and filtered (0.8 mm pore size Millex filter unit, Millipore). The recombinant histidine-tagged proteins were purified by batch purification using 50 ml Ni–NTA agarose beads (QIAGEN) equilibrated with resuspension buffer. Following 3 washes (resuspension buffer), the bound proteins were eluted with resuspension buffer containing 500 mM imidazole. The protein samples were dialyzed at room temperature against 40 mM HEPES (pH 7.0) and 500 mM KCl prior to performing activity assays.

Synthetic psiRNAs

The 45- and 39-nucleotide psiRNAs were chemically synthesized (Integrated DNA Technologies). The sequence of the 45-nucleotide psiRNA 7.01 is: AUUGAAAGUUGUAGUAUGCGGUCCUUGCGGCUGAGAGCACUUCAG. The sequence of the 39-nucleotide psiRNA 7.01 is: AUUGAAAGUUGUAGUAUGCGGUCCUUGCGGCUGAGAGCA.

REFERENCES

1. Makarova, K. S.; Grishin, N. V.; Shabalina, S. A.; Wolf, Y. I.; Koonin, E. V., A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **2006,** 1, 7.

2. Mojica, F. J.; Diez-Villasenor, C.; Garcia-Martinez, J.; Soria, E., Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **2005,** 60, (2), 174-82.

3. Pourcel, C.; Salvignol, G.; Vergnaud, G., CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **2005,** 151, (Pt 3), 653-63.

4. Godde, J. S.; Bickerton, A., The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **2006,** 62, (6), 718-29.

5. Lillestol, R. K.; Redder, P.; Garrett, R. A.; Brugger, K., A putative viral defence mechanism in archaeal cells. *Archaea* **2006,** 2, (1), 59-72.

6. Sorek, R.; Kunin, V.; Hugenholtz, P., CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **2008,** 6, (3), 181-6.

7. Tyson, G. W.; Banfield, J. F., Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **2008,** 10, (1), 200-7.

8. Haft, D. H.; Selengut, J.; Mongodin, E. F.; Nelson, K. E., A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **2005,** 1, (6), e60.

9. Jansen, R.; Embden, J. D.; Gaastra, W.; Schouls, L. M., Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **2002,** 43, (6), 1565-75.

10. Barrangou, R.; Fremaux, C.; Deveau, H.; Richards, M.; Boyaval, P.; Moineau, S.; Romero, D. A.; Horvath, P., CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **2007,** 315, (5819), 1709-12.

11. Deveau, H.; Barrangou, R.; Garneau, J. E.; Labonte, J.; Fremaux, C.; Boyaval, P.; Romero, D. A.; Horvath, P.; Moineau, S., Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol* **2008,** 190, (4), 1390-400.

12. Brouns, S. J.; Jore, M. M.; Lundgren, M.; Westra, E. R.; Slijkhuis, R. J.; Snijders, A. P.; Dickman, M. J.; Makarova, K. S.; Koonin, E. V.; van der Oost, J., Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **2008,** 321, (5891), 960-4.

13. Andersson, A. F.; Banfield, J. F., Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **2008,** 320, (5879), 1047-50.

14. Heidelberg, J. F.; Nelson, W. C.; Schoenfeld, T.; Bhaya, D., Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS One* **2009,** 4, (1), e4169.

15. Tang, T. H.; Bachellerie, J. P.; Rozhdestvensky, T.; Bortolin, M. L.; Huber, H.; Drungowski, M.; Elge, T.; Brosius, J.; Huttenhofer, A., Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc Natl Acad Sci U S A* **2002,** 99, (11), 7536-41.

16. Tang, T. H.; Polacek, N.; Zywicki, M.; Huber, H.; Brugger, K.; Garrett, R.; Bachellerie, J. P.; Huttenhofer, A., Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol Microbiol* **2005,** 55, (2), 469-81.

17. Hale, C.; Kleppe, K.; Terns, R. M.; Terns, M. P., Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. *RNA* **2008,** 14, (12), 2572-9.

18. Lillestol, R. K.; Shah, S. A.; Brugger, K.; Redder, P.; Phan, H.; Christiansen, J.; Garrett, R. A., CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol Microbiol* **2009,** 72, (1), 259-72.

19. Carte, J.; Wang, R.; Li, H.; Terns, R. M.; Terns, M. P., Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* **2008,** 22, (24), 3489-96.

20. Makarova, K. S.; Aravind, L.; Grishin, N. V.; Rogozin, I. B.; Koonin, E. V., A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **2002,** 30, (2), 482-96.

21. Wiedenheft, B.; Zhou, K.; Jinek, M.; Coyle, S. M.; Ma, W.; Doudna, J. A., Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **2009,** 17, (6), 904-12.

22. Beloglazova, N.; Brown, G.; Zimmerman, M. D.; Proudfoot, M.; Makarova, K. S.; Kudritska, M.; Kochinyan, S.; Wang, S.; Chruszcz, M.; Minor, W.; Koonin, E. V.; Edwards, A. M.; Savchenko, A.; Yakunin, A. F., A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem* **2008,** 283, (29), 20361-71.

23. Marraffini, L. A.; Sontheimer, E. J., CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **2008,** 322, (5909), 1843-5.

24. Kunin, V.; Sorek, R.; Hugenholtz, P., Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **2007,** 8, (4), R61.

25. Liu, J.; Carmell, M. A.; Rivas, F. V.; Marsden, C. G.; Thomson, J. M.; Song, J. J.; Hammond, S. M.; Joshua-Tor, L.; Hannon, G. J., Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **2004,** 305, (5689), 1437-41.

26. Elbashir, S. M.; Harborth, J.; Lendeckel, W.; Yalcin, A.; Weber, K.; Tuschl, T., Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **2001,** 411, (6836), 494-8.

27. Elbashir, S. M.; Martinez, J.; Patkaniowska, A.; Lendeckel, W.; Tuschl, T., Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J* **2001,** 20, (23), 6877-88.

28. Schwarz, D. S.; Tomari, Y.; Zamore, P. D., The RNA-induced silencing complex is a Mg2+-dependent endonuclease. *Curr Biol* **2004,** 14, (9), 787-91.

29. Martinez, J.; Tuschl, T., RISC is a 5' phosphomonoester-producing RNA endonuclease. *Genes Dev* **2004,** 18, (9), 975-80.

30. Ghildiyal, M.; Zamore, P. D., Small silencing RNAs: an expanding universe. *Nat Rev Genet* **2009,** 10, (2), 94-108.

31. Malone, C. D.; Hannon, G. J., Small RNAs as guardians of the genome. *Cell* **2009,** 136, (4), 656-68.

32. Chapman, E. J.; Carrington, J. C., Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet* **2007,** 8, (11), 884-96.

33. Zaratiegui, M.; Irvine, D. V.; Martienssen, R. A., Noncoding RNAs and gene silencing. *Cell* **2007,** 128, (4), 763-76.

34. Farazi, T. A.; Juranek, S. A.; Tuschl, T., The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **2008,** 135, (7), 1201-14.

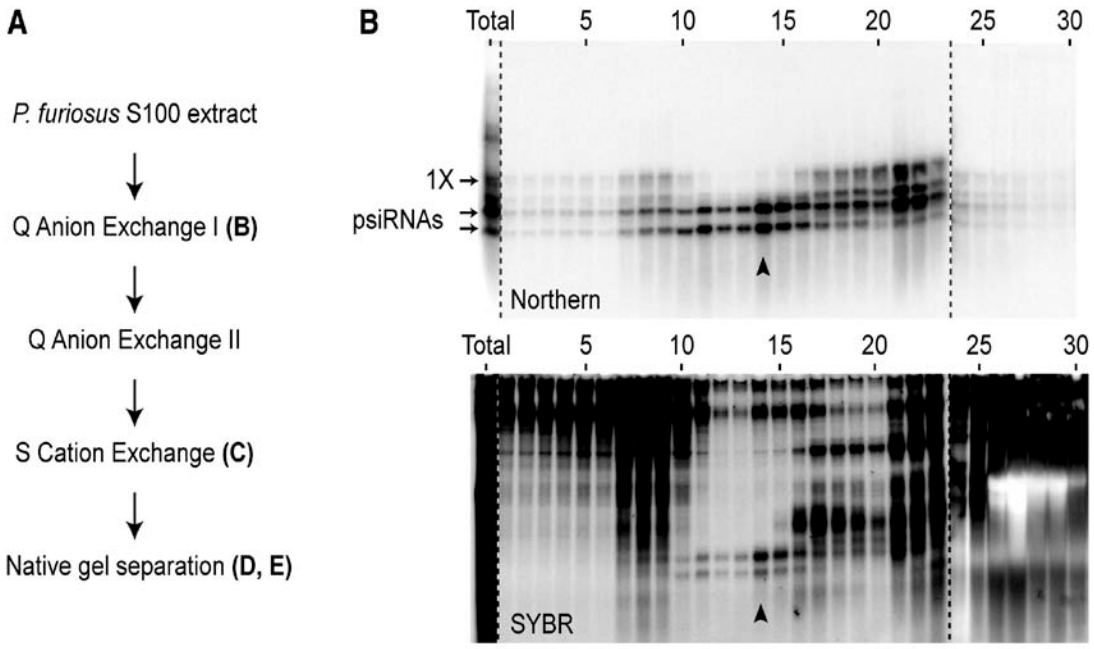
35. Hutvagner, G.; Simard, M. J., Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol* **2008,** 9, (1), 22-32.

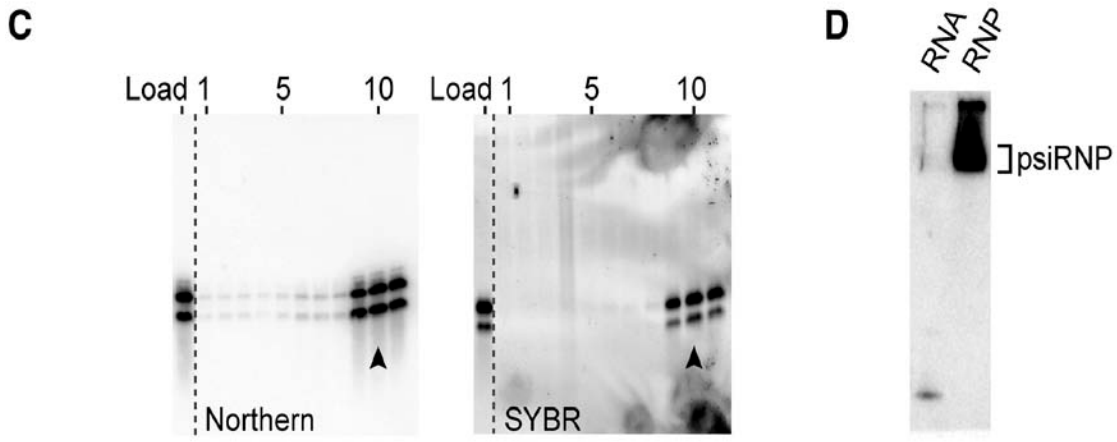
36. Hale, C. R.; Zhao, P.; Olson, S.; Duff, M. O.; Graveley, B. R.; Wells, L.; Terns, R. M.; Terns, M. P., RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* **2009,** 139, (5), 945-956.

37. Wells, L.; Vosseller, K.; Cole, R. N.; Cronshaw, J. M.; Matunis, M. J.; Hart, G. W., Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. *Mol Cell Proteomics* **2002,** 1, (10), 791-804.

38. Lim, J. M.; Sherling, D.; Teo, C. F.; Hausman, D. B.; Lin, D.; Wells, L., Defining the regulated secreted proteome of rodent adipocytes upon the induction of insulin resistance. *J Proteome Res* **2008,** 7, (3), 1251-63.

39. Weatherly, D. B.; Atwood, J. A., 3rd; Minning, T. A.; Cavola, C.; Tarleton, R. L.; Orlando, R., A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics* **2005,** 4, (6), 762-72.


A
P. furiosus S100 extract
Q Anion Exchange I (B)
Q Anion Exchange II
S Cation Exchange (C)
Native gel separation (D, E)
B
Total
5
10
15
20
25
30
1X
psiRNAs
Northern
Total
5
10
15
20
25
30
SYBR


C
Load 1 5 10
Load 1 5 10
Northern
SYBR
D
RNA
RNP
psiRNP

E

Protein	Unique Peptides	Coverage	COG	Gene name	Proposed Function
PF1129	54	55.9%	1353	*cmr2*	Polymerase/Nuclease
PF1128	13	40.4%	1769	*cmr3*	RNA-binding protein (RAMP)
PF1126	14	58.3%	1336	*cmr4*	RNA-binding protein (RAMP)
PF1124	5	16.2%	1604	*cmr6*	RNA-binding protein (RAMP)
PF1125	5	24.9%	3337	*cmr5*	ND
PF0352	5	28.1%	1367	*cmr1-2*	RNA-binding protein (RAMP)
PF1130	1	3.6%	1367	*cmr1-1*	RNA-binding protein (RAMP)

F

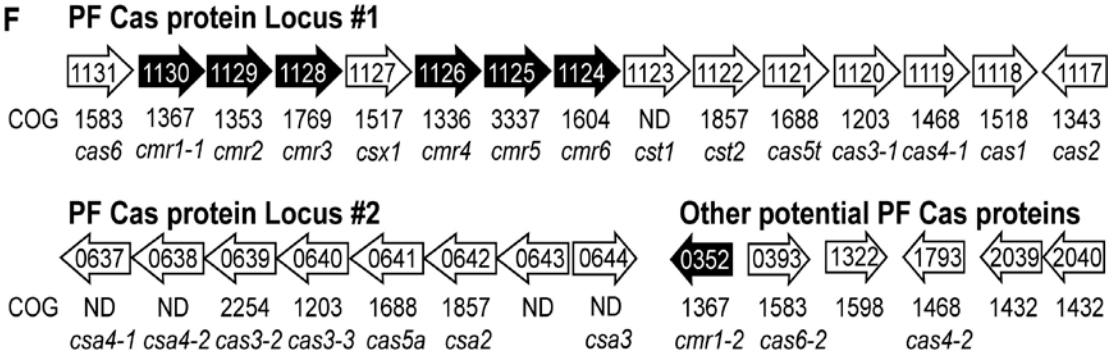
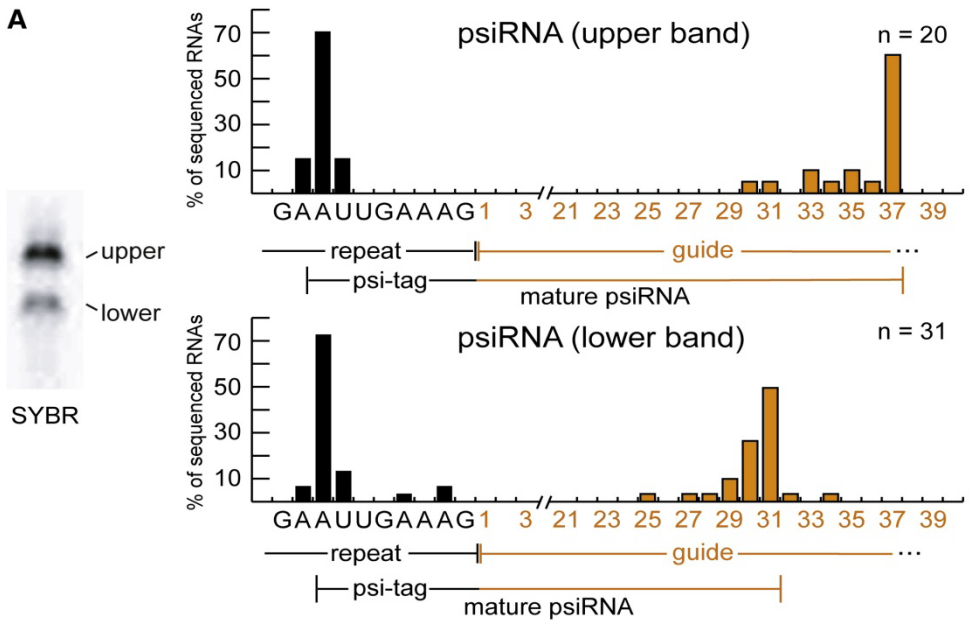


Figure A-1. Identification of a Ribonucleoprotein Complex Containing psiRNAs and Cas Proteins.

(A) psiRNP purification scheme. Letters indicate the location of corresponding data within Figure A-1. **(B)** psiRNA (Northern panel) and total RNA (SYBR panel) profiles across the initial Q-sepharose anion exchange fractions and an unfractionated sample (total). Northern analysis (top panel) was performed for P. furiosus psiRNA 7.01. The positions of the mature psiRNAs and 1X intermediate RNA [17] are indicated. The SYBR panel shows all RNAs detected by SYBR Gold staining. The peak fraction is indicated by an arrow in each panel. Noncontiguous lanes from the same gel (total sample) and a second gel (fractions 24–30) are indicated by dashed lines. **(C)** psiRNA (Northern analysis of psiRNA 7.01) and total RNA (SYBR staining) profiles across the S-sepharose cation exchange fractions and starting material (load). The peak fraction is indicated by an arrow in each panel. Noncontiguous lanes from the same gel are indicated by dashed lines. **(D)** Native gel Northern analysis of the psiRNP. The peak S-sepharose fraction (arrow, [C]) was fractionated by native gel electrophoresis and analyzed by Northern blotting for psiRNA 7.01. RNA extracted from the same fraction was coanalyzed. The position of the psiRNP is indicated. **(E)** Cas proteins identified by tandem mass spectrometry. The isolated psiRNP (D) was subject to in-gel trypsin digestion and tandem mass spectrometry. Sequence coverage and the number of unique peptides for Cas proteins identified with 99% confidence are shown. P. furiosus cas gene names are as given [8] and proposed functions are as predicted [1, 8]. See also Table A-S1. **(F)** Genome organization of predicted P. furiosus cas genes. Operon organization and COG assignments were adapted from NCBI database. Core cas genes (cas) and Cas module-RAMP (cmr), Cas subtype Apern (csa) and Cas subtype Tneap (cst) genes are indicated. Proteins identified by mass spectrometry are indicated in black.


A
upper
lower
SYBR
psiRNA (upper band)
n = 20
% of sequenced RNAs
70
50
30
10
GAAUUGAAAG
1 3 21 23 25 27 29 31 33 35 37 39
repeat
guide
psi-tag
mature psiRNA
psiRNA (lower band)
n = 31
% of sequenced RNAs
70
50
30
10
GAAUUGAAAG
1 3 21 23 25 27 29 31 33 35 37 39
repeat
guide
psi-tag
mature psiRNA

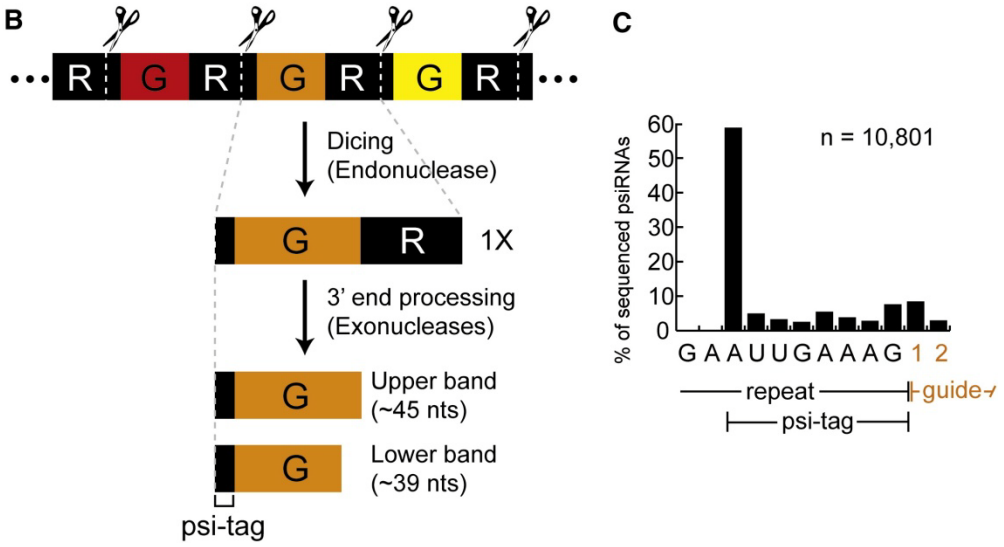
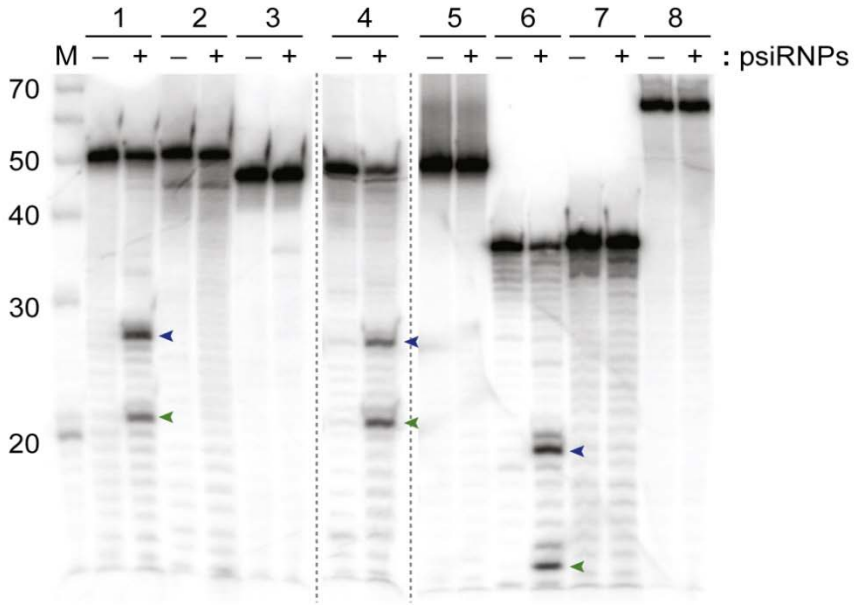


Figure A-2. psiRNA Species in the RNP Contain a Common 5′ Sequence Element and Distinct 3′ Termini.

(A) Sequence analysis of RNAs associated with the complex. RNA species present in the S-sepharose fraction (visualized by SYBR Gold staining) are shown in SYBR panel. RNAs in the upper and lower bands were isolated, cloned, and sequenced. Graphs show the percentage of sequenced RNAs with 5′ ends located at specific positions within the repeat sequence (black), and with indicated numbers of guide sequence nucleotides downstream of the repeat sequence (orange). The average guide sequence is 37 nucleotides in P. furiosus. A consensus for each psiRNA species is diagrammed under each graph. The 8-nucleotide repeat sequence found at the 5′ end of the majority of the psiRNAs is indicated as the psi-tag. **(B)** Model for biogenesis of the two psiRNA species in P. furiosus. CRISPR locus transcripts containing alternating repeat (R, black segments) and guide (G, colored segments) elements are cleaved at a specific site within the repeat by the Cas6 endoribonuclease [19], ultimately producing 1X intermediate RNAs that contain a full invader-targeting sequence flanked on both sides by segments of the repeat. The mature RNAs retain the 5′ end repeat sequence (psi -tag). Uncharacterized 3′ end processing of the 1X intermediate by endo- and/or exo-nucleases forms the two major mature psiRNAs: a 45-nucleotide species that contains the 8-nucleotide psi-tag and a full guide sequence, and a 39-nucleotide species that contains a shorter 31-nucleotide guide sequence. **(C)** Deep sequencing of small RNAs from P. furiosus confirms the presence of the psi-tag. The 5′ ends of the sequenced psiRNAs are graphed as in (A). The number of total clones analyzed (n) is indicated in the graphs of panels (A) and (C).

A

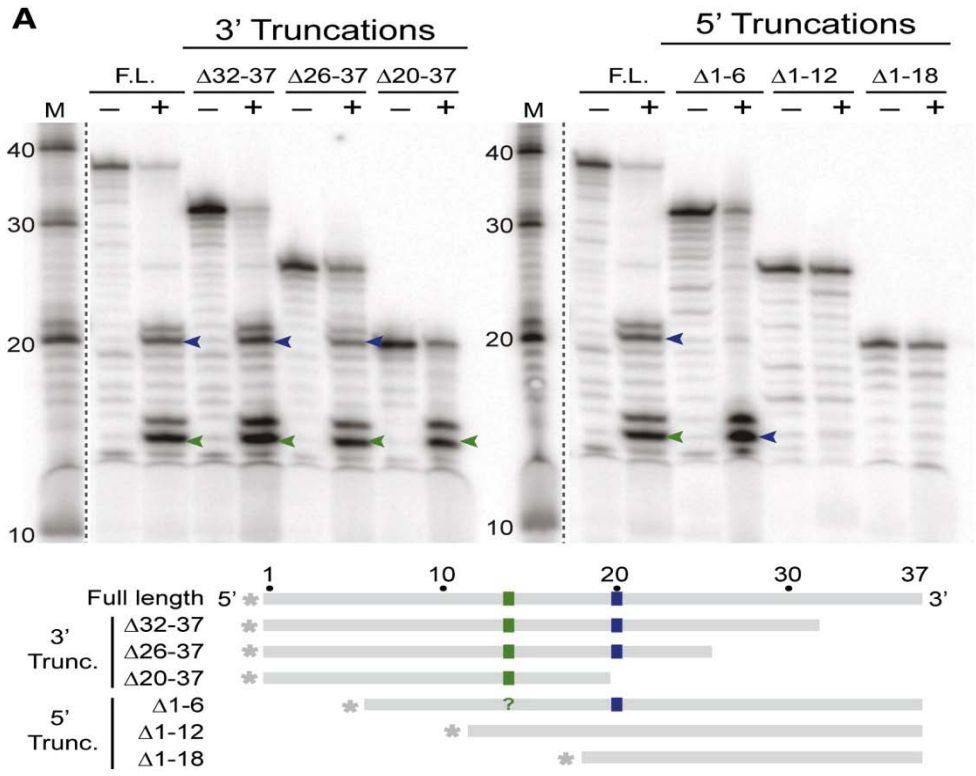


B

#	Substrate	Sequence (5' - 3')
1	7.01 target + ext	AAAAAAA CUGAAGUGCUCUCA | GCCGCA | AGGACCGCAUACUACAA AAAAAA
2	Reverse 7.01 target + ext	AAAAAAA AACAUCAUACGCCAGGAACGCCGACUCUCGUGAAGUC AAAAAA
3	DNA target + ext	AAAAAAA CTGAAGUGCTCTCAGCCGCAAGGACCGCATACTACAA AAAAAA
4	6.01 target + ext	AAAAAAA GUUCCACUAAGGAC | AUUUGU | ACGUCAAAUUCUUCACU AAAAAA
5	Annealed 7.01 target + ext	3′ GACUUCACGAGAGUCGGCGUUCCUGGCGUAUGAUGUU 5′ AAAAAAA CUGAAGUGCUCUCAGCCGCAAGGACCGCAUACUACAA AAAAAA
6	7.01 target	CUGAAGUGCUCUCA | GCCGCA | AGGACCGCAUACUACAA
7	Antisense 7.01 target	UUGUAGUAUGCGGUCCUUGCGGCUGAGAGCACUUCAG
8	sR2 C/D RNA	GGGGATGATGAGTTTTTCCCTCACTCTGATTAGTGATGAGGAGCCGATGCACTGACC

Site 1 Site 2

Figure A-3. Specific Cleavage of Complementary Target RNAs.
The indicated 5′ end-labeled substrates were incubated in the presence (+) or absence (–) of the native psiRNPs (Figure A-1C). Products were resolved by denaturing gel electrophoresis. The primary cleavage products are indicated by green and blue arrows in panel A, and the corresponding sites of cleavage are indicated with green (site 1) and blue (site 2) vertical lines in the substrate sequences shown in panel (B). Noncontiguous lanes from the same gel are indicated by dashed lines, and the sizes of RNA markers (M) are indicated in panel A. "Target" substrates (1, 3, 4, 5, 6) contain regions of perfect complementarity to the guide sequence of the indicated P. furiosus psiRNA. Grey bars demarcate the guide sequences in the panel B. "+ ext" substrates (1, 2, 3, 4, 5) contain 5 and 3′ polyA extensions. For substrate 5, a synthetic psiRNA (sequence shown in gray) was pre-annealed to the 7.01 target RNA + ext. Substrate 2 is a reverse target sequence substrate and substrate 7 is an antisense target substrate. Substrate 3 is DNA; all other substrates are RNA. Substrate 8 is unrelated RNA sR2


A
3' Truncations
5' Truncations
F.L.
Δ32-37
Δ26-37
Δ20-37
F.L.
Δ1-6
Δ1-12
Δ1-18
M
40
30
20
10
1
10
20
30
37
Full length
5'
3'
3' Trunc.
Δ32-37
Δ26-37
Δ20-37
5' Trunc.
Δ1-6
Δ1-12
Δ1-18
?

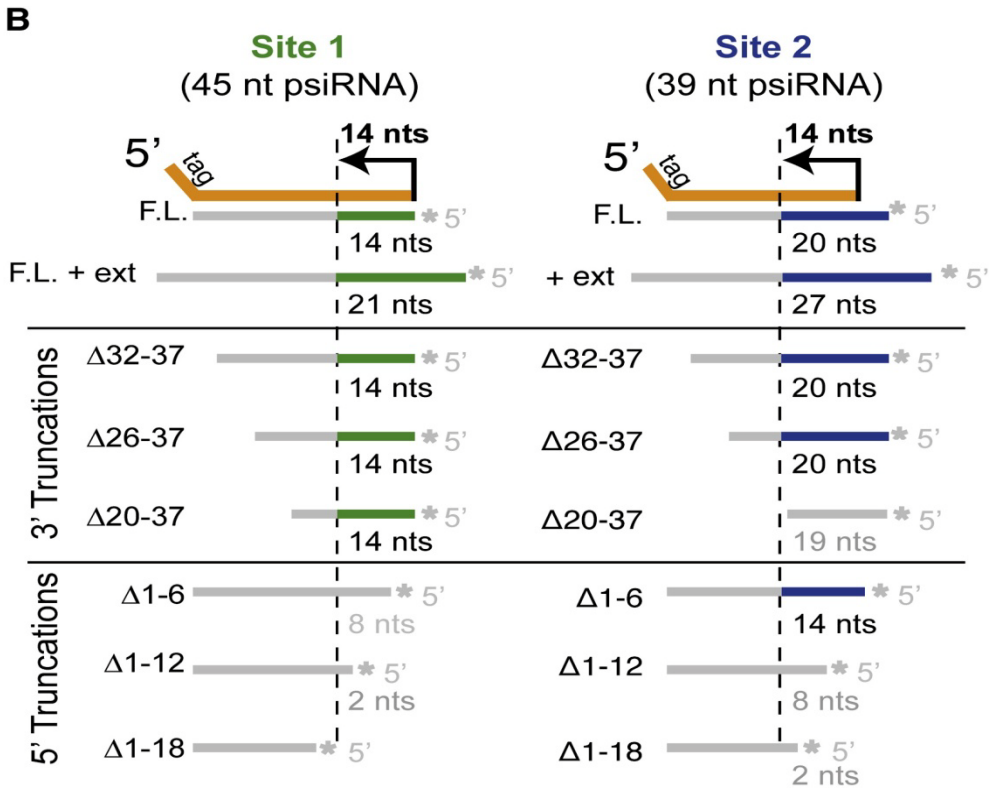
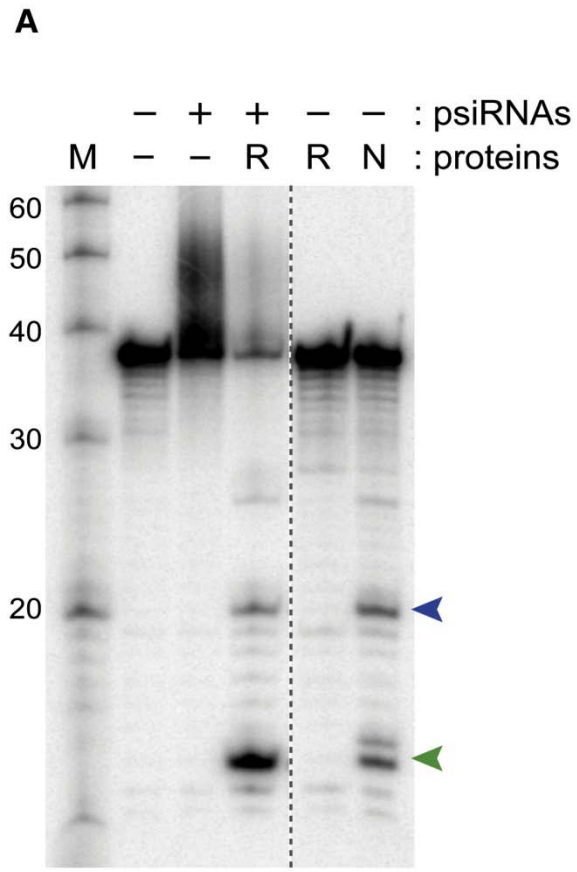


Figure A-4. Cleavage Occurs 14 Nucleotides from the 3′ Ends of the psiRNAs.
(A) The indicated 5′ end -labeled (*) substrates were incubated in the presence (+) or absence (–) of the psiRNP (Figure A-1C). The substrates were full-length 7.01 target RNA (F.L.), and the indicated truncations, and are diagramed below the gel. As in Figure 3, the locations of observed cleavages at sites 1 (green) and 2 (blue) are indicated on the diagrams of the substrate RNAs and the corresponding cleavage products are indicated with green and blue arrows on the gel. The question mark on the diagram of the Δ1-6 target RNA indicates that this cleavage could not be assessed. Noncontiguous lanes from the same gel are indicated by dashed lines. **(B)** Model for cleavage at two sites directed by two psiRNAs. The 45-nucleotide psiRNA species guides cleavage at site 1 and the 39 nucleotide psiRNA guides cleavage at site 2 on each of the substrate RNAs as indicated. In both cases, cleavage occurs 14 nucleotides from the 3′ end of the psiRNA. Observed products are shown in green (site 1) and blue (site 2) and correspond to products in Figures A-3 and A-4A.


A
– + + – – : psiRNAs
M – – R R N : proteins
60
50
40
30
20

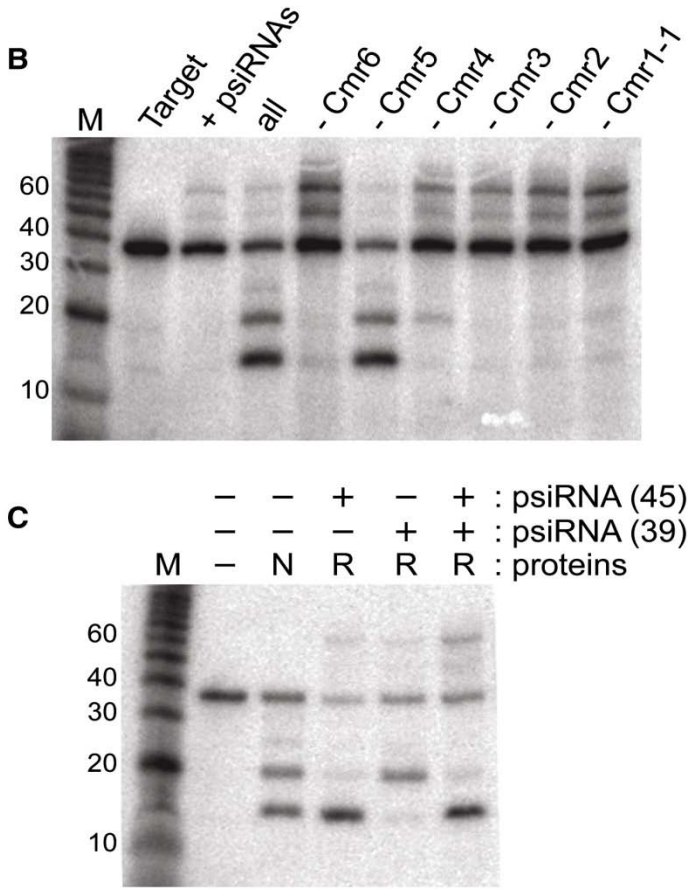


Figure A-5. Target RNA Cleavage Requires Five Cmr Proteins and a Single psiRNA Species.

(A) 5′ end-labeled 7.01 target RNA was incubated in the absence of added psiRNAs or proteins (−), in the presence of synthetic psiRNAs (+) or purified recombinant P. furiosus Cmr proteins (R), or in the presence of purified native psiRNPs (N) as indicated. The synthetic psiRNAs were 45- and 39-nucleotide forms of psiRNA 7.01. The six added recombinant Cmr proteins were P. furiosus Cmr1-1, Cmr2, Cmr3, Cmr4, Cmr5, and Cmr6. Products were resolved by denaturing gel electrophoresis. The products corresponding to cleavage at site 1 and site 2 (see Figure A-3) are indicated by green and blue arrows, respectively. Noncontiguous lanes from the same gel are indicated by a dashed line. The sizes of RNA markers (M) are indicated. **(B)** The 7.01 target RNA (Target) was incubated with the synthetic 7.01 psiRNAs (both 45- and 39-nucleotide species) in the absence (+ psiRNAs) and presence of the purified recombinant P. furiosus Cmr proteins (all), and also with combinations of proteins lacking individual Cmr proteins as indicated (e.g., - Cmr6). **(C)** Cleavage activity of the recombinant psiRNP (R) reconstituted with either the individual 7.01 psiRNA species (45- or 39-nt) or both. Cleavage by the native psiRNP (N) is included for comparison.

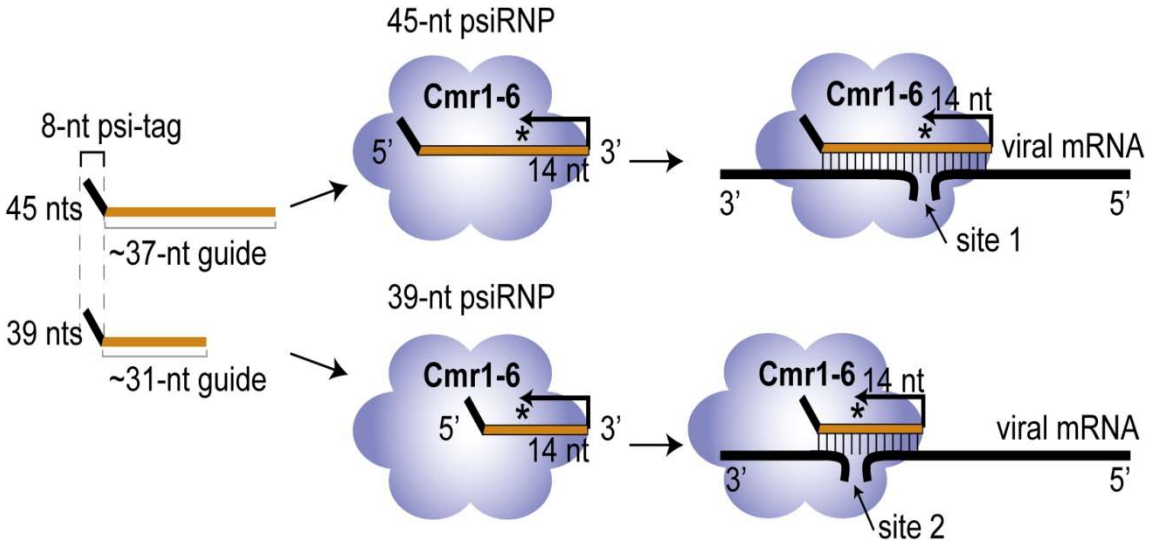


Figure A-6. Model for the Function of psiRNA-Cmr Protein Complexes in Silencing Molecular Invaders.

Based on the results of this study, a psiRNA with a conserved 5′ sequence element derived from the CRISPR repeat (psi-tag) and a region of invader-targeting sequence assembles with six Cas module-RAMP proteins (Cmr1–6). The assembled psiRNP interacts with an invader RNA (e.g., viral mRNA) through base pairing between the psiRNA and invader RNA, positioning the region of the RNA-RNA duplex 14 nucleotides from the 3′ end of the psiRNA in proximity to the active site (star) of the enzyme. In P. furiosus, there are two prominent size forms of each psiRNA with different 3′ ends that guide cleavage of viral mRNAs at two distinct sites. There are also two Cmr1 proteins in P. furiosus that are both found in purified preparations and likely function redundantly.

Table A-S1. Proteins identified by tandem mass spectrometry of native gel-purified RNA protein complexes, expanded from Figure A-1E. All proteins that were identified in the gel-purified complex shown in Figure A-1D (Preparation 1) and in the native gel-purified complex obtained by an alternate chromatography scheme (Preparation 2; see experimental procedures) are listed. Numbers represent the % amino acid sequence coverage, with the number of unique peptides in parentheses. The annotated function is from the NCBI database.

	Protein	Prep 1	Prep 2	Annotated Function
Cas proteins	PF1129	55.9 (54)	26.9 (20)	hypothetical protein PF1129
	PF1128	40.4 (13)	20.5 (5)	hypothetical protein PF1128
	PF1126	58.3 (14)	17.6 (3)	hypothetical protein PF1126
	PF1124	16.2 (5)	11.8 (4)	hypothetical protein PF1124
	PF0352	28.1 (5)		hypothetical protein PF0352
	PF1125	24.9 (5)		hypothetical protein PF1125
	PF1130	3.6 (1)		hypothetical protein PF1130
Non-Cas proteins	PF1717	76.2 (28)		translation initiation factor IF-2 gamma subunit
	PF1683	73.6 (19)		N-acetyl-gamma-glutamyl-phosphate reductase
	PF0990	60.6 (26)		phenylalanyl-tRNA synthetase beta subunit
	PF1685	59.0 (20)		acetylornithine/acetyl-lysine aminotransferase
	PF0481	55.7 (7)		translation initiation factor IF-2 beta subunit
	PF1827	53.8 (14)	20.8 (4)	hypothetical protein PF1827
	PF1881	51.6 (4)		chromatin protein
	PF0989	45.1 (22)		phenylalanyl-tRNA synthetase alpha subunit
	PF0124	34.3 (14)		hypothetical protein PF0124
	PF1140	30.2 (7)		translation initiation factor IF-2 alpha subunit
	PF0495	29.7 (34)		reverse gyrase
	PF1204	29.2 (11)		seryl-tRNA synthetase
	PF1264	26.1 (3)		translation initiation factor IF-5A
	PF0351	25.6 (8)		hypothetical protein PF0351
	PF1238	23.9 (14)		putative ABC transporter
	PF1615	23.1 (18)		hypothetical protein PF1615

Table A-S1. Continued.

	PF0496	21.2 (5)		hypothetical protein PF0496
	PF0594	18.4 (4)	14.3 (2)	Ornithine carbamoyltransferase
	PF1405	16.6 (10)	12.9 (7)	Cleavage and polyadenylation specifity factor protein
	PF0547	15.8 (5)		hypothetical protein PF0547
	PF0969	14.7 (4)		2-ketovalerate ferredoxin oxidoreductase subunit alpha
	PF0220	14.1 (6)	13.6 (7)	Hexulose-6-phosphate synthase
	PF1375	13.3 (6)		elongation factor Tu
	PF1976	12.1 (6)		L-aspartate oxidase
	PF0666	11.5 (6)		nol1-nop2-sun family putative nucleolar protein IV
	PF1746	11.0 (6)		hypothetical protein PF1746
	PF0251	11.0 (4)		hypothetical protein PF0251
	PF1579	10.4 (7)		DNA topoisomerase VI subunit B
	PF0966	10.4 (4)		2-oxoglutarate ferredoxin oxidoreductase
	PF0533	10.1 (7)		indolepyruvate ferredoxin oxidoreductase subunit a
	PF1578	9.2 (3)		DNA topoisomerase VI subunit A
	PF0026	8.8 (4)		tRNA nucleotidyltransferase
	PF1540	8.7 (5)		ADP forming acetyl coenzyme A synthetase
	PF1203	8.1 (5)		formaldehyde:ferredoxin oxidoreductase
	PF1046	7.9 (3)		queuine trna-ribosyltransferase
	PF0464	7.5 (4)		glyceraldehyde-3-phosphate:ferredoxin oxidoreductase
	PF1768	5.1 (2)		2-oxoglutarate ferredoxin oxidoreductase
	PF0440	3.9 (5)		ribonucleotide-diphosphate reductase alpha subunit
	PF1843	1.7 (2)	7.3 (6)	chromosome segregation protein smc
	PF0102		76.6 (15)	hypothetical protein PF0102
	PF1883		74.9 (13)	small heat shock protein
	PF1548		63.3 (24)	hypothetical protein PF1548
	PF1931		27.9 (6)	hypothetical protein PF1931
	PF0162		11.2 (2)	hypothetical protein PF0162
	PF0204		6.0 (2)	hypothetical protein PF0204
	PF1871		3.7 (1)	"N(2),N(2)-dimethylguanosine tRNA methyltransferase"
	PF1245		2.2 (1)	hypothetical d-nopaline dehydrogenase
	PF1167		1.5 (1)	chromosome segregation protein

APPENDIX B

MOUSE AND ZEBRAFISH HOXA3 ORTHOLOGUES HAVE NONEQUIVALENT IN VIVO PROTEIN FUNCTION[1]

CONTRIBUTION: Used offline strong cation exchange chromatography to separate peptides from mouse embryos, and LC-MS/MS method to quantify the target peptide (SPLLNSPTVGK).

[1]Reprinted with permission from the National Academy of Sciences
Lizhen Chen, Peng Zhao, Lance Wells, Chris T. Amemiya, Brian G. Condie, Nancy R. Manley
Proc. Natl. Acad. Sci. U. S. A., 107, 23, 10555-60, 2010.

ABSTRACT

Hox genes play evolutionarily conserved roles in specifying axial position during embryogenesis. A prevailing paradigm is that changes in Hox gene expression drive evolution of metazoan body plans. Conservation of Hox function across species, and among paralogous Hox genes within a species, supports a model of functional equivalence. In this report, we demonstrate that zebrafish hoxa3a (zfhoxa3a) expressed from the mouse Hoxa3 locus can substitute for mouse Hoxa3 in some tissues, but has distinct or null phenotypes in others. We further show, by using an allele encoding a chimeric protein, that this difference maps primarily to the zfhoxa3a C-terminal domain. Our data imply that the mouse and zebrafish proteins have diverged considerably since their last common ancestor, and that the major difference between them resides in the C-terminal domain. Our data further show that Hox protein function can evolve independently in different cell types or for specific functions. The inability of zfhoxa3a to perform all of the normal roles of mouse Hoxa3 illustrates that Hox orthologues are not always functionally interchangeable.

INTRODUCTION

Hox genes encode a family of transcription factors with conserved roles in patterning the anterior–posterior axis during embryogenesis in all bilaterian animals [1]. Mice and other mammals have 39 Hox genes arranged in four clusters located on four different chromosomes, whereas zebrafish and other teleosts have 48 Hox genes in seven clusters resulting from a genome-wide duplication [2]. Hox genes from the same group (transparalogous genes or paralogues) arose from duplication, and share more similarity in protein sequence and expression pattern than genes within a cluster. Paralogous Hox genes often play diverse biological roles, as evidenced by their mutant phenotypes, but also show extensive redundancy and functional overlap. When vertebrate Hox genes have been expressed in Drosophila, they often provide similar functions as their Drosophila orthologues, supporting a model of functional equivalence of Hox genes cross phyla. In contrast to this striking functional conservation during evolution, many studies have shown a correlation between changes in Hox expression pattern and variation in morphological pattern [3-4]. These results have lead to a widely accepted model that cis-element evolution is the main driving force of morphological evolution, and is a major mechanism whereby Hox genes participate in this process [5]. Although specific instances of protein functional divergence have been correlated with morphological evolution in arthropods [6], the degree to which cis-regulatory versus protein function changes influence morphological evolution remains controversial [5, 7]. Furthermore, the largely non-segmented body plan in vertebrates and increased potential for redundancy as a result of extra genome duplications raises the question of which mechanism(s) are operational in vertebrates.

The group 3 Hox genes are required for patterning part of the anterior body plan during embryogenesis. Extensive genetic studies of mouse Hoxa3 have demonstrated roles in patterning and development of endodermal, mesodermal, and ectodermal derivatives, and in cell migration, proliferation, apoptosis, and differentiation. Hoxa3 is expressed in the third and fourth pharyngeal pouch endoderm and in pharyngeal arch mesenchyme, and has similar anterior boundaries in multiple tissues [8]. Null mutation of mouse Hoxa3 causes neonatal lethality, pharyngeal organ defects or aplasia, and defects in the tracheal epithelium, soft palate, pharyngeal skeleton, the IX cranial nerve, and the carotid body [8-12]. Hoxa3 mutation also exacerbates the defects of single or compound mutants of Group 3 Hox paralogues in the axial skeleton and neural tube [10, 13-14].

The striking functional equivalence of Hox3 paralogues was most dramatically demonstrated by swapping the Hoxa3 and Hoxd3 protein coding sequences, leaving the regulatory regions intact [15]. Expressing either protein under the control of the other's regulatory sequences resulted in a WT phenotype, providing strong evidence that these two proteins are functionally equivalent despite their different single mutant phenotypes and diverged protein sequences. Hoxa1 and Hoxb1 were also largely interchangeable by a similar approach [16]. These and other studies support the functional equivalence of paralogous Hox proteins, and suggest that the overall quantity of Hox proteins may be more important than the specific proteins present [17-18].

In this study we tested the conservation of Hoxa3 orthologues between mouse and zebrafish, the two major vertebrate model organisms. To test whether the zebrafish hoxa3a protein can substitute for mouse Hoxa3, we generated an allele in which mouse Hoxa3 coding sequence was precisely replaced with that of zfhoxa3a, the only Hoxa3

orthologue in zebrafish [19]. The zebrafish gene expressed from mouse locus complemented some defects seen in the mouse Hoxa3 null mutant, consistent with the equivalence of mouse Hoxa3 and Hoxd3, showing that all three proteins share conserved biological functions. However, the Hoxa3zf allele was equivalent to the null in the development of the IXth cranial nerve, thymus, and parathyroid, and had a neomorphic pharyngeal skeleton phenotype. Using a second strain in which only the C-terminal half of the protein is from zebrafish, we show that these functional differences primarily map to the domain downstream of the homeodomain. Although protein sequence alignment showed that overall, the zebrafish hoxa3a and mouse Hoxa3 and Hoxd3 proteins showed similar degrees of conservation, zebrafish hoxa3a appears to have undergone extremely rapid molecular evolution relative to other vertebrate Hoxa3 orthologues. These data provide evidence that the zebrafish hoxa3a and mouse Hoxa3 proteins have functionally diverged since their respective taxa last shared a common ancestor, and suggest that these differences are to the results of changes outside the homeodomain.

RESULTS

Expression of Zebrafish hoxa3a Protein from Mouse Hoxa3 Locus

To generate a new Hoxa3 allele (Hoxa3zf) that expressed the zfhoxa3a protein from the endogenous mouse Hoxa3 locus, the mouse Hoxa3 protein coding sequences were replaced with those of zebrafish hoxa3a by gene targeting, and a C-terminal HA tag was added, similar to the strategy used previously for the mouse Hoxa3-Hoxd3 swap (Figure B-1A) [15]. All sequences outside the protein coding domains, including the intron between the two coding exons, were from the mouse locus. Another allele, Hoxa3mz, was produced as a consequence of recombination occurring within the mouse intron and

in the homologous sequences 3′ of the neoR cassette (Figure B-1A). Hoxa3mz encodes a protein with mouse N-terminal domain (NTD) and hexapeptide sequences and zebrafish homeodomain and C-terminal domain (CTD).

As recent data have identified transcription factor binding sites within the coding sequence of the Hoxa2 gene [20], we tested whether the zf and mz alleles had the same mRNA expression patterns and levels as the WT mouse allele. At embryonic d 10.5, the Hoxa3zf allele was expressed correctly, with the same spatial and temporal pattern and at the same level as the WT Hoxa3 mRNA (Figure B-1B-1F). Analysis of the zebrafish hoxa3a protein using the HA tag showed that the protein was present and had the correct anterior limit in the hindbrain (Figure B-1G-1I).

Conserved Hoxa3 Protein Functions

We tested whether zebrafish hoxa3a protein was able to substitute for mouse Hoxa3 in mice that expressed only the zf or mz alleles (zf/zf or mz/mz; Table B-1), or in compound heterozygotes with the null allele (zf/null or mz/null). In the Hoxa3-null mutant mouse, the ventral thyroid isthmus is absent or ectopic (Figure B-2A and 2B), and the ultimobranchial body–derived C-cells fail to mix with the thyroid proper (Figure B-2E and 2F) [8]. Neither of these phenotypes were present in Hoxa3zf/zf and Hoxa3mz/mz mice (Figure B-2C, 2D, 2G, and 2H). The tracheal epithelium in Hoxa3-null mutants has a thicker epithelial layer and a convoluted surface [8] (Figure B-2I and 2J). In all the Hoxa3zf/zf and Hoxa3mz/mz animals examined, this epithelium was normal (Figure B-2K and 2L). These phenotypes were also rescued by only one copy of the Hoxa3zf or Hoxa3mz allele (zf/null and mz/null).

Hoxa3null/null animals have a truncated secondary palate [9] that can result in a bloated abdomen caused by breathing air into the esophagus (Figure B-2M and 2N). In all Hoxa3zf/zf, Hoxa3mz/mz, and Hoxa3mz/null newborns analyzed, the secondary palate was normal, and the bloated abdomen phenotype was never present (Figure B-2O and 2P). The exception was in zf/null mice, which had only a partial rescue of palate length that was unable to prevent bloating. Thus, the zf and mz alleles were able to completely substitute for mouse Hoxa3 in the thyroid/ultimobranchial body and tracheal epithelium, and in most cases in the soft palate. Interestingly, mutants of all genotypes (zf/zf, mz/mz, zf/null, mz/null) died within hours after birth, similar to the null mutant [9], indicating that other phenotypes contribute to the neonatal lethality of Hoxa3 null mutants.

Diverged Hoxa3 Protein Functions

Although zebrafish hoxa3a was sufficient for the normal development of some tissues, it was equivalent to a null allele in other aspects of the phenotype. In the majority of Hoxa3-null mutants, the IXth cranial nerve is either fused to the vagus nerve (X) th or disconnected from the hind-brain [10, 12]. Similar defects at similar frequencies were observed in Hoxa3zf/zf and Hoxa3mz/mz embryos (Figure B-3A-3D, Table B-1).

The most consistent phenotype of the Hoxa3-null mutant is the absence of thymus and parathyroids. Neither organ was detected at the normal or any ectopic location in Hoxa3zf/zf and Hoxa3mz/mz animals (Figure B-3E-3L). These morphological defects were supported by earlier changes in marker expression associated with thymus and parathyroid organogenesis. The thymus-specific marker, Foxn1 [21], was absent from the third pharyngeal pouch in both Hoxa3null/null and Hoxa3zf/zf E11.5 embryos.

Expression of Gcm2, which is required for parathyroid organogenesis [22-23], was also greatly reduced in both mutants. Like the null mutant, the expression of Pax1, a potential downstream target of Hoxa3, was also reduced specifically in the third pouch of the Hoxa3zf/zf embryos at embryonic d 10.5 [8] (Figure B-3M-3O).

These results show that both Hoxa3zf and Hoxa3mz act as null alleles in cranial nerve IX, thymus, and parathyroid development. As the homeodomain is identical between the mouse and zebrafish proteins, these functional differences map to the downstream CTD.

Morphological Differences in Hyoid Development from zf and mz Alleles

The lesser and greater horns of the hyoid bone are derived from the second and third pharyngeal arches, respectively. In the Hoxa3null/null mouse, the lesser horn is absent or greatly reduced and the greater horn is malformed and fused to the thyroid cartilage [10, 14] (Figure B-4B). The Hoxa3zf/zf and Hoxa3mz/mz mutants had similar greater horn phenotypes as the null mutant, with fusions to the thyroid cartilage, which was also malformed (Figure B-4C and 4D, black arrows). Although both these mutants did rescue the presence of a lesser horn, in both cases it had morphologies different from WT. The zf/zf lesser horn had a teardrop shape that was fused to the middorsal cranial aspect of the greater horn (Figure B-4C, white arrow). The lesser horn in the mz/mz mutant was more square than WT or zf/zf, and there was an extra cartilage extension from the greater horn back to the base of the skull that was not seen in the WT or zf/zf genotypes (Figure B-4D, white arrow). Neither phenotype was affected by reducing the dose in zf/null or mz/null mice. These data suggest that hyoid lesser horn patterning is sensitive to a species difference in Hoxa3 protein function, and that both the NTD and

CTD may contribute. These phenotypes were always recessive to the WT morphology, as heterozygotes for either the zf or mz alleles had normal morphologies, showing that the mouse protein was dominant to the zebrafish protein in establishing pharyngeal skeletal morphology.

Quantitative Analysis of Zebrafish-Derived Proteins in Mice

Although the zf and mz alleles were correctly expressed at the mRNA level, it was possible that their failure to fully substitute for the mouse protein was a result of reduced translation or protein stability. We tested whether the mouse and zebrafish proteins had similar steady-state protein levels in vivo using a proteomics approach to quantify protein levels using MS. We used the mz allele for this test, as it shared the NTD with the mouse WT protein and is functionally similar to the zf allele, and measured the amount of each protein in whole mz/mz homozygous or WT E10.5 embryos. Using a tryptic peptide from the mouse Hoxa3 NTD that was not found in other Hox proteins, the normalized area under the peak from a reconstructed ion chromatogram was compared between the two samples. The ratio of Hoxa3 between mz/mz and WT mice was calculated as 1.2:1; this slight difference was not significant. The similar levels and localization of mRNA and protein for the mouse and zebrafish proteins show that any functional differences between these alleles are a result of functional differences between these proteins.

Interaction Between zfHoxa3 and Hoxd3 in the Axial Skeleton

Although Hoxa3null/null mutants have a normal axial skeleton, reducing Hoxa3 dosage in a Hoxd3 mutant background reveals a dosage-dependent functional redundancy for these genes [10, 14]. Reducing the dose of Hoxa3 in Hoxd3 homozygotes leads to

progressively more severe defects of the atlas and basioccipital bone, with the entire atlas deleted in Hoxa3−/−;Hoxd3−/− mice (Figure B-4E-4H) [14]. Hoxa3zf/zf;Hoxd3−/− mice had the same phenotype as Hoxa3−/−;Hoxd3−/− (Figure B-4G and 4J), suggesting that Hoxa3zf allele might function as a null allele in the cervical vertebrae. However the zf allele did not show the same dose dependency as the mouse null, as Hoxa3+/zf;Hoxd3−/− mice had a phenotype more similar to Hoxa3+/+;Hoxd3−/− than Hoxa3+/−; Hoxd3−/− (Figure B-4I). These results suggest that zebrafish hoxa3a has a function intermediate to the null and WT alleles of mouse Hoxa3 in patterning the cervical vertebrae.

zfhoxa3a Failed to Substitute for Mouse Hoxa3 in Neural Crest

To test the tissue specificity of zfhoxa3-associated phenotypes, we generated Wnt1cre;Hoxa3fx/zf animals, in which NCCs expressed only zfhoxa3a, but all other tissues were heterozygous (+/zf). These mice had phenotypes essentially identical to a NCC-specific KO of Hoxa3 (Figure B-5). Deletion of a conditional allele of Hoxa3 in neural crest cells (NCCs) using Wnt1cre caused defects in thymus and parathyroid morphogenesis, including failure to separate from the pharynx and delayed detachment of the parathyroids in Wnt1cretg/0;Hoxa3fx/null mice (Figure B-5A, 5B, 5D, and 5E). Thymic lobes were ectopic and still connected to the pharynx (Figure B-5C), with overall size and organization of these lobes similar to the WT thymus. The parathyroid was ectopic but normal in morphology (Figure B-5F-5H). Thus, zebrafish hoxa3a failed to substitute for mouse Hoxa3 for its function in NCCs to support later stages of thymus and parathyroid morphogenesis.

Accelerated Evolution of the zfhoxa3a Protein

To further map the sequences responsible for these phenotypic differences, we compared the protein sequences of mouse Hoxa3, Hoxd3, and zebrafish hoxa3a genes. As Hoxa3 and Hoxd3 are functionally similar, their comparison acts as a baseline for estimating similarity based on primary protein sequence. The overall identity between zfhoxa3a and muHoxa3 (59%) was slightly higher than that between muHoxa3 and muHoxd3 (52%). Murine Hoxa3 and zfhoxa3a have identical hexapeptide and homeodomain sequences, and the region between these domains was also highly conserved; in these DNA binding and cofactor interaction domains, the mouse and zebrafish Hoxa3 proteins were more similar than were the mouse Hoxa3 and Hoxd3 proteins. The region N-terminal to the hexapeptide contained a proline-rich domain in mouse Hoxa3 that was absent in mouse Hoxd3 and zfhoxa3a. The CTDs of the three proteins shared higher homology than the N-terminal regions, and had several short regions that were highly conserved in all three proteins. Overall, there were no obvious regions of high homology between the two mouse proteins that were absent or diverged in the zebrafish protein.

As a more sensitive test of sequence divergence, we performed molecular evolutionary analyses of select Hoxa3 proteins (without homeodomain) from cartilaginous fishes, bony fishes, and various tetrapods. A neighbor-joining tree rooted to the cartilaginous fish lineage showed that the zfhoxa3a sequence is considerably distant from other vertebrate Hoxa3 sequences, including those of other teleost fishes. This suggests that its rate of molecular evolution may be accelerated relative to other vertebrates. Relative rate tests (RRTs) using the Tajima method [24] confirmed that the

zebrafish gene has undergone a significantly faster rate of evolution than orthologues from all other vertebrate taxa, including bony fishes. Interestingly, although functional differences primarily mapped to the CTD, we did not observe overt differences in tree topology or relative rates of evolution between the NTD versus CTD of Hoxa3 when the respective sequences were stratified.

DISCUSSION

Cross-species functional tests have provided valuable information in understanding the evolution of gene function, including Hox genes. To our knowledge, the present study is the first cross-species study to use precise gene replacement in mice as previously used to show functional homology of paralogous mouse Hox proteins. We have used this approach to test the functional equivalence of the Hoxa3 gene from two distantly related vertebrate species. Mouse Hoxa3 is required for the development of a diverse range of structures and tissues, providing a unique opportunity to test Hox gene function. The application of quantitative proteomics technology to measure steady-state protein levels provides a high degree of confidence that the zebrafish-derived mRNA sequences are correctly transcribed and translated, and that the resulting fish-derived proteins have similar steady-state levels as the mouse protein. The combination of this genetic approach with the ability to assay a wide range of phenotypic parameters allowed us to test with great sensitivity whether zebrafish hoxa3a is equivalent to mouse Hoxa3 when they are expressed in the same biological context.

Nonequivalence of Mouse and Zebrafish Hoxa3 Proteins

Our data show that the mouse and zebrafish Hoxa3 proteins have nonequivalent in vivo functions. This result is not predicted by the paradigm of functional equivalence,

which holds that conserved transcriptional regulators, such as Hox proteins, are generally equivalent between paralogues and orthologues. The use of genetic complementarity studies across species was an early aspect of vertebrate Hox gene studies, and the ability of human, mouse, and chicken Hox genes to rescue Drosophila Hox mutations or to mimic overexpression phenotypes revealed a high degree of functional equivalence between distant Hox orthologues [25-27]. The ability of vertebrate Hox genes to function in an insect indicates that vertebrate Hox proteins have retained many ancestral functions across long evolutionary distances and dramatic morphological changes. However, it is possible that Hox proteins may have also evolved novel functions during vertebrate evolution, in addition to retaining ancient functions.

Our data suggest that, unlike the mouse paralogous Hox3 genes, Hoxa3 protein function has evolved considerably since the divergence of the ancestral species that gave rise to mammals and teleosts. This result is surprising, as the genome-wide duplication that generated the Hoxa3 and Hoxd3 paralogs occurred at least 450 Mya, and before the divergence of teleost fishes from the vertebrate lineage leading to mammals [2]. As a result, zebrafish hoxa3a diverged from mouse Hoxa3 more recently than mouse Hoxd3, but shows clear evidence of rapid molecular evolution in both the NTD and CTD. However, only the CTD showed evidence of functional divergence. The combination of molecular evolution analysis with the genetic test of functional equivalence provides strong evidence that the zebrafish hoxa3a is functionally distinct. The inability of zfhoxa3a to perform all of the functions of Hoxa3 shows that orthologues are not always functionally equivalent, and shows that Hox protein divergence appears to occur in vertebrates, as has also been shown in arthropods [6]. Analysis of the coelacanth genome compared with

mammalian and teleost genomes, and particularly of Hox gene clusters, indicates that teleosts have considerable genetic plasticity relative to the lineage leading to mammals, with an extra genome duplication and rapid genetic and phenotypic radiation [28-30]. Our data suggest that the zebrafish hoxa3a gene has evolved at both the gene and protein function level compared with its mammalian orthologs. Although the additional genome duplication in zebrafish and subsequent subfunctionalization could account for individual gene differences in some cases, in this case the duplicated gene (i.e., Hoxa3b) has not been retained and therefore cannot be providing the missing functions, further supporting our conclusion of protein functional divergence.

Several molecular mechanisms could explain the failure of zfhoxa3a to substitute fully for mouse Hoxa3. Hoxa3 and hoxa3a may have independently evolved novel functions in addition to ancestral protein functions and/or lost ancestral functions since the divergence of these two vertebrate lineages. Our evidence for rapid molecular evolution of the zfhoxa3a gene suggests that the zebrafish protein may have acquired or lost functions compared with the mouse. It is also possible that within each species, each protein is required for the same functions, but that the genetic networks within which they function have coevolved such that zfhoxa3a cannot interact properly with components of the mouse network in some cell types. It is intriguing in this respect that only some functions are lost whereas others are retained, suggesting that Hoxa3 may interact with independent gene networks in different cell types or at different times during development.

CTD of the Hoxa3 Protein

Our data also show that the majority of the functional differences between mouse and zebrafish Hoxa3 proteins maps to the second coding exon, which includes the homeodomain and CTD. Although differences in homeodomain sequence in different mouse paralogous Hox groups can be associated with functional differences [31-32], zebrafish hoxa3a has an identical homeodomain to mouse Hoxa3. Thus, the functional differences that we have identified between these two Hox proteins must reside downstream of the homeodomain. Our data may provide the first in vivo evidence for specific required functions for vertebrate Hox proteins outside of the conserved hexapeptide and homeodomain regions. Interestingly, only group 2 and group 3 Hox proteins have a long CTD, which is present in group 3 Hox proteins at least as far back as tunicates [33]. Therefore, this domain could be a critical target for the functional evolution of Hox3 proteins. CTD sequences that are conserved between mouse Hoxa3 and Hoxd3 proteins, but diverged in zebrafish hoxa3a, are thus candidates for mediating this functional divergence.

Hox Protein Function and Morphological Evolution

Hoxa3 was the first Hox gene to be mutated in mice [9], and the diversity of phenotypes in Hoxa3-null mutants has allowed us to analyze the functional equivalence of the mouse and zebrafish orthologues in much greater detail than has been possible in prior studies of crossspecies functional conservation [8-10, 13-14]. Given the different morphologies in the pharyngeal regions of fish and mammals and the hypothesized role of Hox genes in morphological evolution, it is tempting to speculate whether the differences that we see between the mouse and zebrafish Hoxa3 orthologues are caused

in part by differing demands of mouse and fish anatomy. Such questions are significantly complicated by the difficulty of unambiguously assigning structural homology in organisms as diverged as mice and fish; however, for some phenotypes such as the lack of a thymus, it is clear that the zebrafish gene is failing to function in the generation of a structure that is present in both species. As a hoxa3a-null zebrafish mutant is not available, we do not know what functions zfhoxa3a performs in zebrafish. There are a total of four Hox3 genes in teleosts, one each for the a3, b3, c3, and d3 groups [19]. Unlike in mouse, at least two and possibly three group 3 Hox genes are expressed in the pouch endoderm in zebrafish. A previous report using morpholino (MO) knockdown suggested that loss of hoxa3a alone had little phenotypic effect, but showed redundant function with Hoxb3a in the development of gill-related structures [34]. However, the phenotypic effects occur at relatively late stages, and it was not clear whether the MOs remained effective. We performed a similar analysis using splice-blocking MOs. Although the MOs were effective at 24 h, by 52 h, spliced mRNAs were readily apparent, indicating that the lack of phenotypes at later stages may not reflect the full range of gene function. Thus, it remains possible that zfhoxa3a has some similar functions to the mouse gene. Alternatively, some aspects of Hoxa3 protein function could be performed by other Hox3 genes in zebrafish. If so, this would represent a novel combination of gene expression and protein function-based subfunctionalization during the evolution of Hox3 genes in vertebrates.

The intriguing result that some functions seem to be shared between the mouse and zebrafish Hoxa3 proteins whereas others are not may have important implications for the ability of “toolkit” transcription factors to evolve at the protein level to effect

morphological change [5]. Hoxa3 is an excellent example of "mosaic pleiotropy," in which a single transcription factor has diverse functions in different structures and at different times. This concept has been proposed as a principal reason why any change in function of toolkit transcription factors would not likely be tolerated [5]. That the ability of the zebrafish protein to substitute for mouse Hoxa3 is not universal, and ranges between completely WT to completely null phenotypes, indicates that Hoxa3 may perform different roles or interact with different partners in different cell types, and that these functions may evolve independently to some extent. Furthermore, these differences map outside of the conserved homeodomain, indicating that subtle changes in less well conserved domains may have major and specific effects on protein function without affecting DNA binding, and thus may serve as important sites of Hox protein evolution.

MATERIALS AND METHODS

Gene Targeting

The mouse Hoxa3zf and Hoxa3mz alleles were generated by homologous recombination. The Hoxa3 locus was targeted with a vector based on a 12-kb Not I fragment of C57Bl6 genomic DNA that was linearized and electroporated into LK-1 C57Bl6 ES cells. ES cell line derivation, electroporation, and injection were performed in the Mouse ES Cell and Transgenic Core Facility at the Medical College of Georgia. Clones were screened by Southern blot with 5′ and 3′ franking probes and an internal probe.

Hoxa3 mRNA and Protein Quantification

First-strand cDNA was reverse transcribed from total RNA from embryonic d 10.5 embryos. Quantitative PCR was performed on an ABI 7500 real-time PCR system

with SYBR green PCR master mix (Applied Biosystems). For proteomics analysis, the target peptide (SPLLNSPTVGK) was chosen from the tryptic peptides of the mouse Hoxa3 NTD. Proteins extracted from mouse embryos were digested with trypsin (Promega) following reduction and alkylation. Resulting peptides were separated with an offline strong cation exchange chromatography. Seven fractions were collected and analyzed in parent ion monitoring mode via LC-MS/MS (LTQ-Orbitrap XL; ThermoFisher). Acquired spectra were searched against a mouse protein database (Swissprot, updated March 24, 2009) using Bioworks (version 3.3.1, SP1; ThermoFisher). The calculation of the ratio was based on the peak area of the reconstructed ion chromatogram of respective peptides following normalization by a high-scoring tryptic peptide that coeluted from titin. Details are provided in SI Materials and Methods.

SUPPORTING INFORMATION MATERIALS AND METHODS

Hoxa3 Relative Protein Quantification

The target peptide (SPLLNSPTVGK) was chosen from the tryptic peptides of mouse Hoxa3 using the criteria of it not being conserved in other Homeobox family proteins and not having amino acids susceptible to oxidation and alkylation during sample workup. Proteins were extracted from mouse embryos (two embryos for each genotype, +/+ and mz/mz) and digested by trypsin (Promega) following reduction and alkylation. The resulting peptides were separated by an offline strong cation exchange chromatography. The tryptic peptides were separated by offline strong cation exchange liquid chromatography. Solvent A (5 mM KH2PO4/30% acetonitrile, pH 2.7), solvent B (solvent A with 350 Mm KCl), and solvent C (0.1 M Tris/0.5 M KCl, pH 7.0) were used

to develop a linear gradient consisting of 5 min at 100%solvent A, 48 min gradient at variable slope to 100% solvent B, 12 min at 100% solvent B, 15 min to 100% solvent C, and 10 min to 100% solvent A. Fractions were collected every 2 min, and then combined into five fractions, desalted, and dried. Seven fractions were collected for each genotype sample, and each fraction was analyzed in SRM mode via LC-MS/MS (LTQ-Orbitrap XL;ThermoFisher). The acquired spectra were searched against a mouse protein database (Swissprot, updated on March 24, 2009) using Bioworks (version3.3.1 SP1; Thermo Fisher Scientific). The calculation of the ratio was based on the peak area of the reconstructed ion chromatogram of respective peptides following normalization by a high-scoring trpytic peptide that coeluted from titin in both samples.

REFERENCES

1. McGinnis, W.; Krumlauf, R., Homeobox genes and axial patterning. *Cell* **1992,** 68, (2), 283-302.

2. Amores, A.; Suzuki, T.; Yan, Y. L.; Pomeroy, J.; Singer, A.; Amemiya, C.; Postlethwait, J. H., Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res* **2004,** 14, (1), 1-10.

3. Wray, G. A., The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **2007,** 8, (3), 206-16.

4. Gellon, G.; McGinnis, W., Shaping animal body plans in development and evolution by modulation of Hox expression patterns. *Bioessays* **1998,** 20, (2), 116-25.

5. Carroll, S. B., Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **2008,** 134, (1), 25-36.

6. Pearson, J. C.; Lemons, D.; McGinnis, W., Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **2005,** 6, (12), 893-904.

7. Lynch, V. J.; Wagner, G. P., Resurrecting the role of transcription factor change in developmental evolution. *Evolution* **2008,** 62, (9), 2131-54.

8. Manley, N. R.; Capecchi, M. R., The role of Hoxa-3 in mouse thymus and thyroid development. *Development* **1995,** 121, (7), 1989-2003.

9. Chisaka, O.; Capecchi, M. R., Regionally restricted developmental defects resulting from targeted disruption of the mouse homeobox gene hox-1.5. *Nature* **1991,** 350, (6318), 473-9.

10. Manley, N. R.; Capecchi, M. R., Hox group 3 paralogous genes act synergistically in the formation of somitic and neural crest-derived structures. *Dev Biol* **1997,** 192, (2), 274-88.

11. Kameda, Y.; Nishimaki, T.; Takeichi, M.; Chisaka, O., Homeobox gene hoxa3 is essential for the formation of the carotid body in the mouse embryos. *Dev Biol* **2002,** 247, (1), 197-209.

12. Watari, N.; Kameda, Y.; Takeichi, M.; Chisaka, O., Hoxa3 regulates integration of glossopharyngeal nerve precursor cells. *Dev Biol* **2001,** 240, (1), 15-31.

13. Manley, N. R.; Capecchi, M. R., Hox group 3 paralogs regulate the development and migration of the thymus, thyroid, and parathyroid glands. *Dev Biol* **1998,** 195, (1), 1-15.

14. Condie, B. G.; Capecchi, M. R., Mice with targeted disruptions in the paralogous genes hoxa-3 and hoxd-3 reveal synergistic interactions. *Nature* **1994,** 370, (6487), 304-7.

15. Greer, J. M.; Puetz, J.; Thomas, K. R.; Capecchi, M. R., Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature* **2000,** 403, (6770), 661-5.

16. Tvrdik, P.; Capecchi, M. R., Reversal of Hox1 gene subfunctionalization in the mouse. *Dev Cell* **2006,** 11, (2), 239-50.

17. Duboule, D., Vertebrate Hox genes and proliferation: an alternative pathway to homeosis? *Curr Opin Genet Dev* **1995,** 5, (4), 525-8.

18. Duboule, D., Developmental genetics. A Hox by any other name. *Nature* **2000,** 403, (6770), 607, 609-10.

19. Amores, A.; Force, A.; Yan, Y. L.; Joly, L.; Amemiya, C.; Fritz, A.; Ho, R. K.; Langeland, J.; Prince, V.; Wang, Y. L.; Westerfield, M.; Ekker, M.; Postlethwait, J. H., Zebrafish hox clusters and vertebrate genome evolution. *Science* **1998,** 282, (5394), 1711-4.

20. Tumpel, S.; Cambronero, F.; Sims, C.; Krumlauf, R.; Wiedemann, L. M., A regulatory module embedded in the coding region of Hoxa2 controls expression in rhombomere 2. *Proc Natl Acad Sci U S A* **2008,** 105, (51), 20077-82.

21. Gordon, J.; Bennett, A. R.; Blackburn, C. C.; Manley, N. R., Gcm2 and Foxn1 mark early parathyroid- and thymus-specific domains in the developing third pharyngeal pouch. *Mech Dev* **2001,** 103, (1-2), 141-3.

22. Gunther, T.; Chen, Z. F.; Kim, J.; Priemel, M.; Rueger, J. M.; Amling, M.; Moseley, J. M.; Martin, T. J.; Anderson, D. J.; Karsenty, G., Genetic ablation of parathyroid glands reveals another source of parathyroid hormone. *Nature* **2000,** 406, (6792), 199-203.

23. Liu, Z.; Yu, S.; Manley, N. R., Gcm2 is required for the differentiation and survival of parathyroid precursor cells in the parathyroid/thymus primordia. *Dev Biol* **2007,** 305, (1), 333-46.

24. Tajima, F., Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **1993,** 135, (2), 599-607.

25. Leuzinger, S.; Hirth, F.; Gerlich, D.; Acampora, D.; Simeone, A.; Gehring, W. J.; Finkelstein, R.; Furukubo-Tokunaga, K.; Reichert, H., Equivalence of the fly orthodenticle gene and the human OTX genes in embryonic brain development of Drosophila. *Development* **1998,** 125, (9), 1703-10.

26. Acampora, D.; Avantaggiato, V.; Tuorto, F.; Barone, P.; Reichert, H.; Finkelstein, R.; Simeone, A., Murine Otx1 and Drosophila otd genes share conserved genetic functions required in invertebrate and vertebrate brain development. *Development* **1998,** 125, (9), 1691-702.

27. Lutz, B.; Lu, H. C.; Eichele, G.; Miller, D.; Kaufman, T. C., Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. *Genes Dev* **1996,** 10, (2), 176-84.

28. Koh, E. G.; Lam, K.; Christoffels, A.; Erdmann, M. V.; Brenner, S.; Venkatesh, B., Hox gene clusters in the Indonesian coelacanth, Latimeria menadoensis. *Proc Natl Acad Sci U S A* **2003,** 100, (3), 1084-8.

29. Crow, K. D.; Stadler, P. F.; Lynch, V. J.; Amemiya, C.; Wagner, G. P., The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Mol Biol Evol* **2006,** 23, (1), 121-36.

30. Noonan, J. P.; Grimwood, J.; Danke, J.; Schmutz, J.; Dickson, M.; Amemiya, C. T.; Myers, R. M., Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res* **2004,** 14, (12), 2397-405.

31. Zhao, Y.; Potter, S. S., Functional specificity of the Hoxa13 homeobox. *Development* **2001,** 128, (16), 3197-207.

32. Zhao, Y.; Potter, S. S., Functional comparison of the Hoxa 4, Hoxa 10, and Hoxa 11 homeoboxes. *Dev Biol* **2002,** 244, (1), 21-36.

33. Pierce, R. J.; Wu, W.; Hirai, H.; Ivens, A.; Murphy, L. D.; Noel, C.; Johnston, D. A.; Artiguenave, F.; Adams, M.; Cornette, J.; Viscogliosi, E.; Capron, M.; Balavoine, G., Evidence for a dispersed Hox gene cluster in the platyhelminth parasite Schistosoma mansoni. *Mol Biol Evol* **2005,** 22, (12), 2491-503.

34. Hogan, B. M.; Hunter, M. P.; Oates, A. C.; Crowhurst, M. O.; Hall, N. E.; Heath, J. K.; Prince, V. E.; Lieschke, G. J., Zebrafish gcm2 is required for gill filament budding from pharyngeal ectoderm. *Dev Biol* **2004,** 276, (2), 508-22.

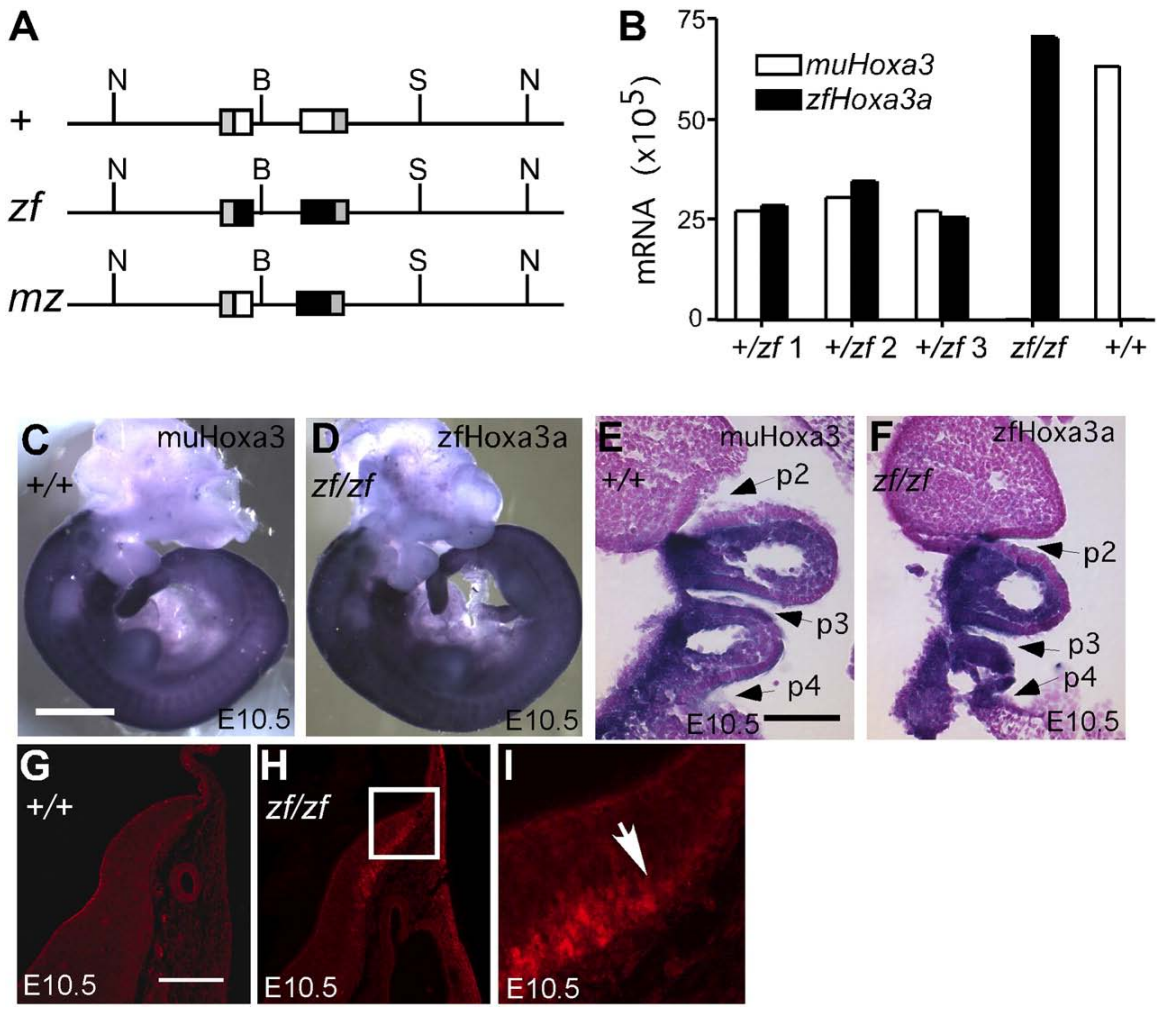


Figure B-1. Structure and expression of Hoxa3 alleles.
(A) Scheme of Hoxa3 WT (+), Hoxa3zf (zf), and Hoxa3mz (mz) alleles. Horizontal thin lines represent noncoding genomic DNA at the mouse Hoxa3 locus. Boxes represent exons as follows: gray, 5′ or 3′UTR of mouse Hoxa3; white, mouse Hoxa3 coding sequences; black, zebrafish hoxa3a coding sequences. N, NotI; B, BamHI; S, SpeI. **(B)** Quantitative RT-PCR shows equivalent mRNA levels for the zfhoxa3a (zf) and mouse Hoxa3 (WT) transcripts in individual whole heterozygous embryos (+/zf 1–3), or in homozygotes (zf/zf, +/+). Whole mount **(C** and **D)** and coronal paraffin section **(E** and **F)** in situ hybridization of embryonic d 10.5 embryos using allele-specific probes shows identical spatial expression patterns for the WT murine and zf alleles. **(G-I)** Immunofluorescence detection of the HA tag in the zfHoxa3 protein in the hindbrain at embryonic d 10.5. Box in H corresponds to panel in I. Arrow in I shows anterior limit of protein detection. (Scale bars: 1 mm in C and D; 200 μm in G-I.)

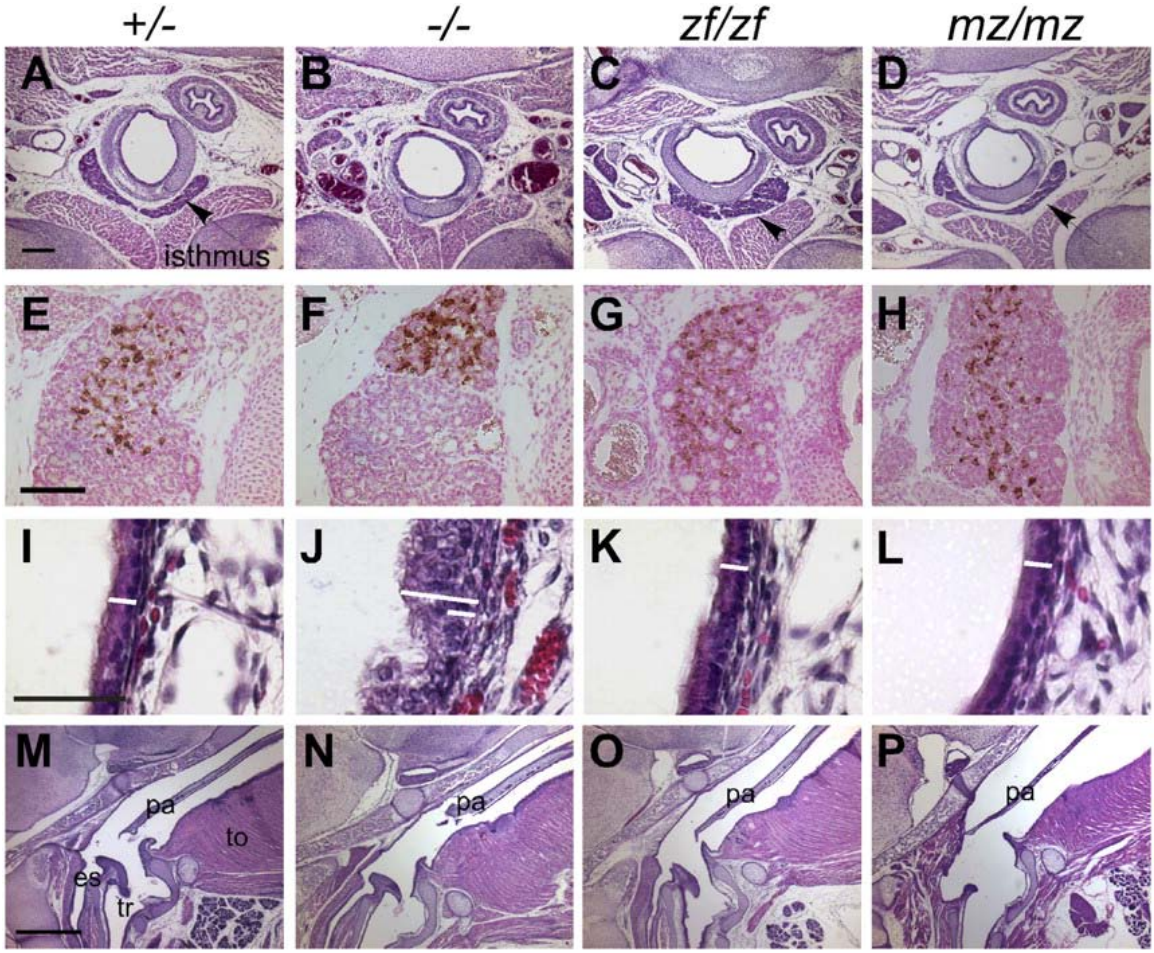


Figure B-2. Zebrafish hoxa3a can substitute for mouse Hoxa3 in thyroid, ultimobranchial body, tracheal epithelium, and soft palate development.
Transverse (**A-L**; dorsal is up) or sagittal (**M-P**; anterior is up, dorsal is to the left.) paraffin sections of newborn animals; genotypes apply to each column. Scale bars apply to each row. **(A-D)** The thyroid isthmus (arrow) is deleted in Hoxa3null/null(−/−) mice (**B**), but restored in Hoxa3zf/zf (zf/zf) (**C**) and Hoxa3mz/mz (mz/mz) (**D**). (**E-H**) Transverse sections of newborn mice stained with anticalcitonin antibody (brown). Integration of ultimobranchial body–derived C cells is restored in zf/zf (**G**) and mz/mz (**H**) mice. (**I-L**) The disorganized tracheal epithelium in the Hoxa3-null mutant (**J**) was not seen in zf/zf (**K**) or mz/mz (**L**) animals. The white bar in each panel shows the thickness of the WT epithelium, contrasted with the null mutant (long bar in J). (**M-P**) The posterior palate (velum) is shortened in Hoxa3-null mutants (**N**), but is normal in zf/zf and mz/mz mice (M, O, and P). tr, trachea; es, esophagus; pa, palate; to, tongue. (Scale bars: 200 μm in A-D; 100 μm in E-H; 50 μm in I–L; 800 μm in M-P.)

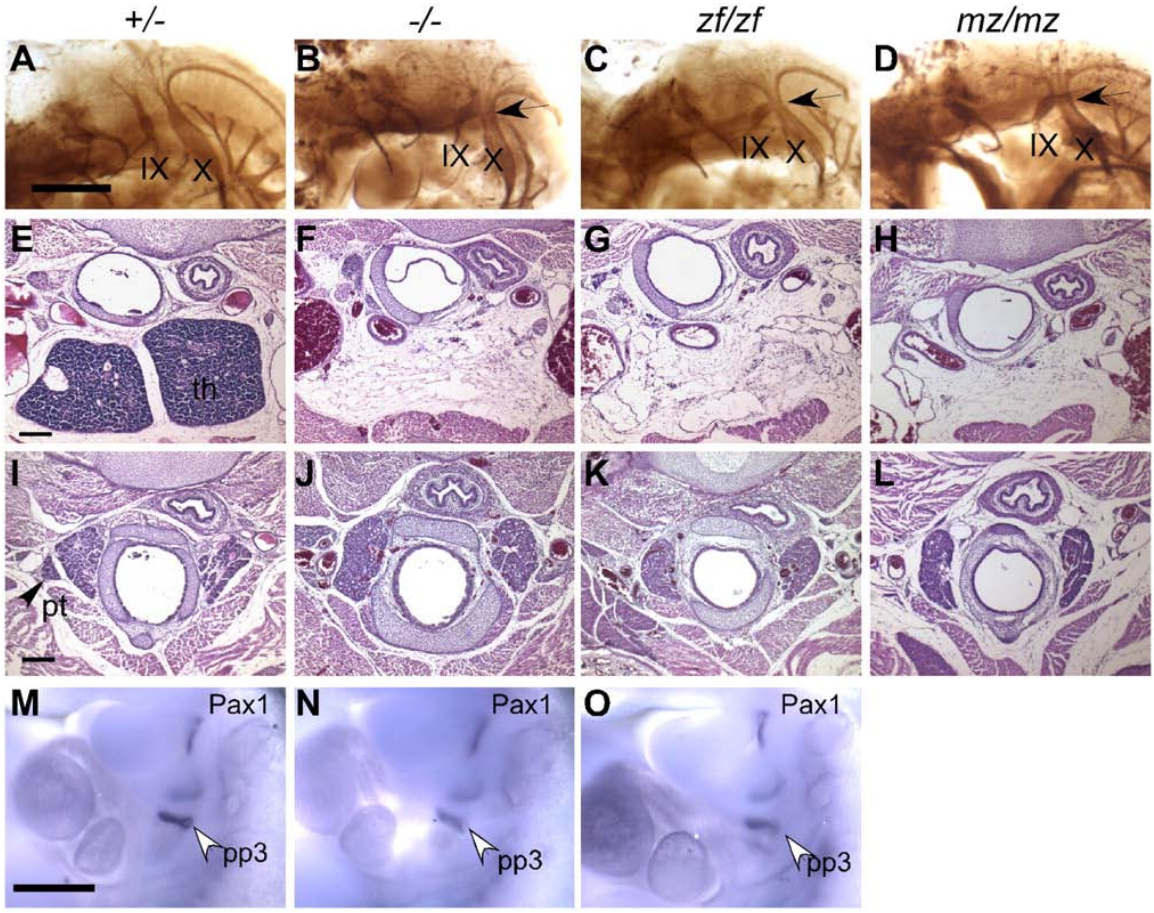


Figure B-3. Cranial nerve, thymus, and parathyroid defects are not rescued by zfhoxa3a.

(**A-D**) Whole-mount antineurofilament staining of embryonic d 10.5 control, Hoxa3null/null (−/−), Hoxa3zf/zf (zf/zf), and Hoxa3mz/mz (mz/mz) embryos. In the control, the IX cranial nerve is connected to hindbrain. (**B**) In −/− embryos, the IX cranial nerve is often fused (arrow) to the X cranial ganglia. (**C** and **D**) The same fusion is observed in zf/zf and mz/mz mutants. (**E-L**) Transverse paraffin sections of newborn animals stained with hematoxylin and eosin (dorsal is up). (Scale bar: 200 μm.) The thymus (**F-H**) and parathyroids (**J-L**) are absent in −/−, zf/zf, and mz/mz mutant mice. th, thymus; pt, parathyroid. (**M-O**) Whole-mount in situ hybridization for Pax1 at embryonic d 10.5. Pax1 expression in the third pouch (pp3) is reduced in −/− embryo, but expression in the other pouches is unchanged. zf/zf shows a similar pattern as −/− (cranial is up). (Scale bar: 500 μm.)

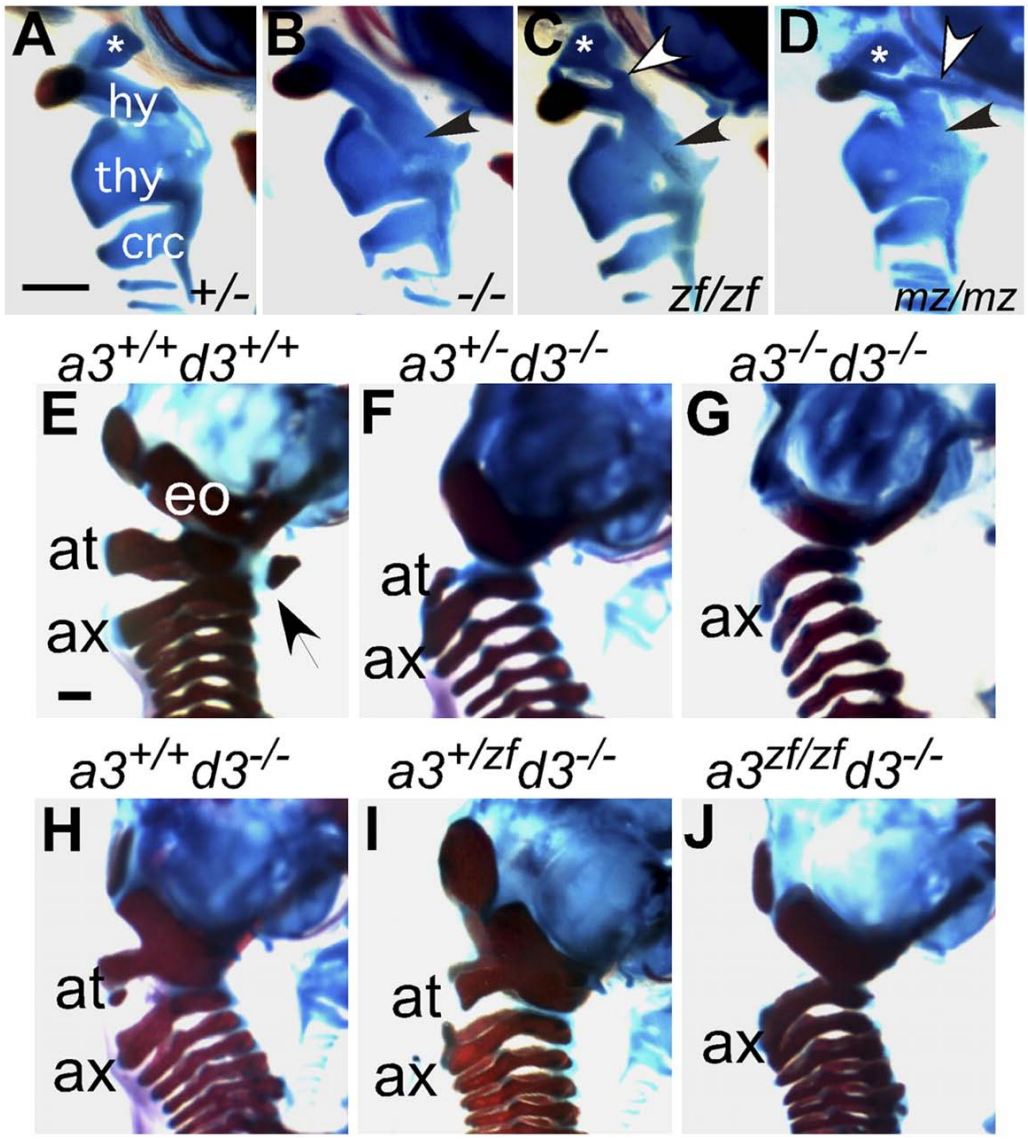


Figure B-4. Novel pharyngeal skeleton morphologies in Hoxa3zf/zf and Hoxa3mz/mz mice, and skeletal phenotype of compound mutants with Hoxd3.
(**A-D**) Lateral views of the throat cartilages in cleared newborn skeletal preparations. Anterior is up, dorsal is to the right. Asterisk indicates the lesser horn of hyoid; hy, greater horn of hyoid; thy, thyroid cartilage; crc, cricoid cartilage. (Scale bar: 500 μm.) In Hoxa3null/null (−/−), Hoxa3zf/zf (zf/zf), and Hoxa3mz/mz (mz/mz), the greater horn is malformed and fused to the thyroid cartilage (black arrows in B-D). (**B**) In the null, the lesser horn of the hyoid is greatly reduced or deleted. (**C** and **D**) The zf/zf and mz/mz mutants have distinct hyoid morphologies, and are different from WT or null. White arrows show extra cartilage structures in these mutants. (**E-J**) Lateral views of the cervical region in cleared skeleton preparations of the indicated genotypes for Hoxa3 (a3) or Hoxd3 (d3). Anterior is up, dorsal is to the left. Exoccipital (eo) bone, atlas (at), axis (ax), and anterior arch of atlas (arrowhead) are indicated. Note that G and J are similar, whereas I is more similar to H than to F. (Scale bar: 1 mm.)

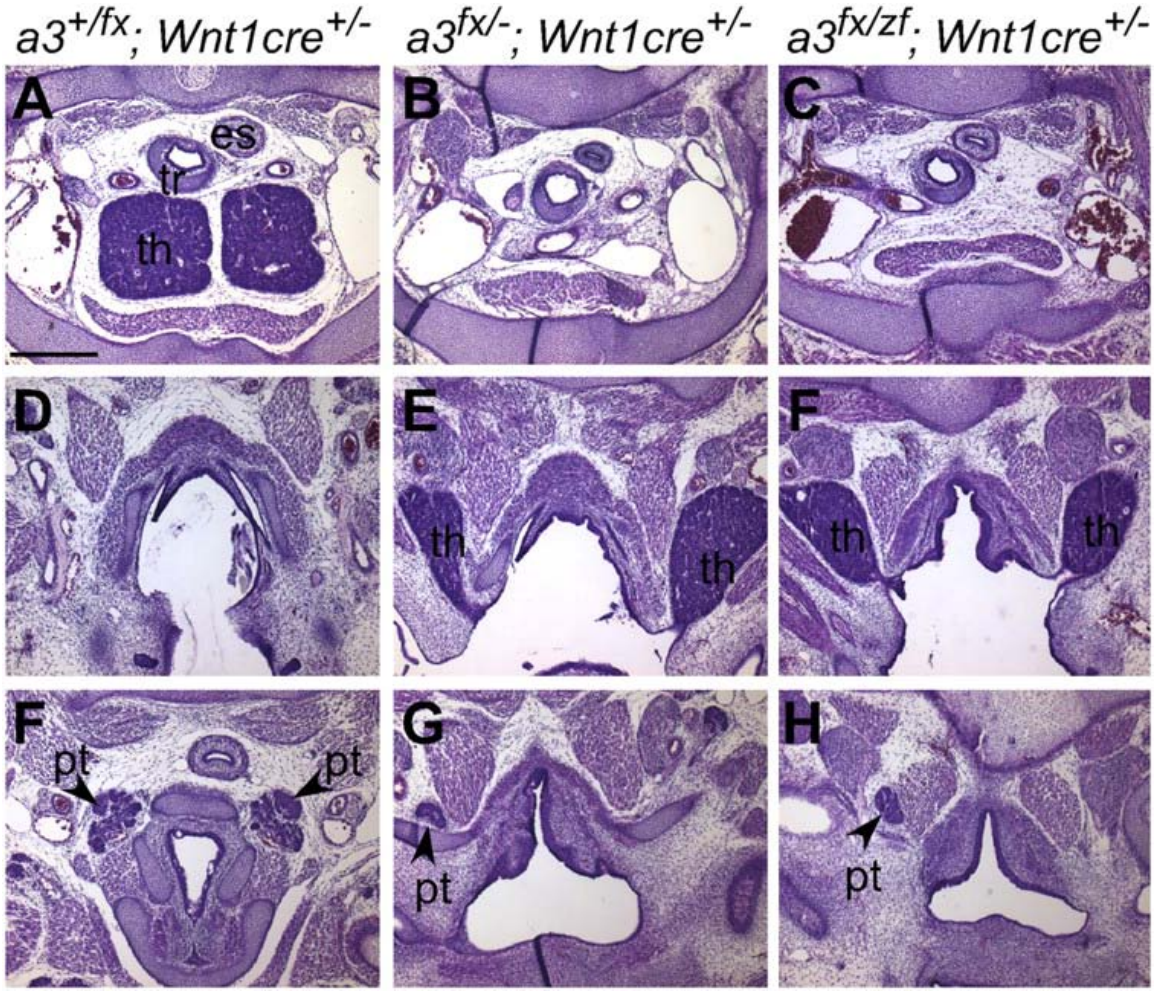


Figure B-5. Hoxa3zf allele has null function in NCCs.

Hematoxylin and eosin staining of transverse paraffin sections from embryonic d 15.5 embryos (dorsal is up). Genotypes apply to each column; panels in each row are from a comparable anterior–posterior location. In the control embryo, the thymus (th) is located in the chest (**A**), and the parathyroids (pt) are embedded in the thyroid (**F**). In embryos with a NCC-specific deletion of mouse Hoxa3 (Hoxa3fx/-;Wnt1cre+/−) the thymic lobes are absent from the normal position (**B**), and are instead still attached to the pharynx and are ectopic (**E**). Parathyroids are also ectopic and anterior to the thymus (**G**). (**C**, **F**, and **H**) Embryos in which only the zf allele is expressed in the NCC (Hoxa3fx/zf;Wnt1cre+/−) have a phenotype identical to NCC-specific Hoxa3 deletion. (Scale bar: 400 μm.)

Table B-1. Summary of the phenotypes, showing that Hoxa3mz allele functions virtually the same as Hoxa3zf allele.

Location	Hoxa3$^{null/null}$	Hoxa3$^{zf/zf}$	Hoxa3$^{mz/mz}$
Thyroid isthmus	Deleted or ectopic	WT	WT
Ultimobranchial body	Separated from thyroid	WT	WT
Tracheal epithelium	Disorganized	WT	WT
Soft palate	Truncated	WT	WT
IX cranial nerve	Disconnected or fused to X	Null	Null
Thymus	Athymia	Null	Null
Parathyroid	Aparathyroidism	Null	Null
Throat cartilage	Malformed	Null	Null
Hyoid lesser horn	Deleted	Neomorphic*	Neomorphic†

*The hyoid lesser horn is different in morphology from WT.
†The lesser horn of Hoxa3$^{mz/mz}$ appears different from either WT or Hoxa3zf/zf.

APPENDIX C

THE OUTER MEMBRANE PROTEOME OF BURKHOLDERIA PSEUDOMALLEI AND BURKHOLDERIA MALLEI FROM DIVERSE GROWTH CONDITIONS: INSIGHT INTO ABUNDANT PROTEINS ALWAYS PRESENT ON OR NEAR THE CELL SURFACE[1]

CONTRIBUTION: Used LC-MS/MS tandem mass spectrometry to identify membrane proteins in *Burkholderia mallei* and *Burkholderia pseudomallei*.

[1] Manuscript submitted to Proteomics
Mark A. Schell, Peng Zhao, Lance Wells

ABSTRACT

Burkholderia mallei and Burkholderia pseudomallei are closely related aerosol-infective human pathogens that can cause life-threatening diseases. Biochemical analyses requiring large scale growth and manipulation at BSL3 under select agent regulations can be cumbersome and hazardous. We developed a simple, safe, and rapid method to prepare highly purified outer membrane (OM) fragments from these pathogens. Shotgun proteomic analyses of these membranes by trypsin shaving and tandem mass spectrometry resulted in identification of 170 proteins, the majority of which are clearly outer membrane proteins (OMPs). These included: 14 porins, multiple secretins for virulence factor export, several efflux pumps, multiple components of a Type VI secreton, TonB-dependent metal transport receptors, polysaccharide exporters, and hypothetical OMPs of unknown function. We also identified 20 OMPs in the OM of each bacterium that are abundantly produced under a wide variety of growth conditions suggesting these are fundamental for growth of these pathogens and thus prime drug or vaccine targets. Comparison of the OM proteomes of Bp and Bm showed many similarities but also revealed some clear differences perhaps reflecting evolution of B. mallei away from environmental survival toward host-adaptation.

INTRODUCTION

Many macromolecules in the outer membrane (OM) of bacterial pathogens, especially proteins exposed on the cell surface, are important virulence factors, as well as targets for host immune recognition. Identification of abundant and/or novel outer membrane proteins (OMPs) and characterization of their roles in pathogen physiology, disease, and defense against the host immune system is an important preliminary step in development of diagnostics, vaccines, and antibacterials. Until recently, OMPs were usually identified individually using two dimensional (2D) gel electrophoresis of solubilized membranes followed by tryptic digestion of individual spots and peptide mass fingerprint identification with mass spectrometry.[1-4] However, recent advances in liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) technology has opened new avenues to identify en masse large numbers of proteins in whole cells or subcellular fractions.[5] The OM is an excellent subcellular fraction to target for en masse shotgun proteomics since its complexity is low, i.e. it contains a relatively small number of protein species in comparison to the cytoplasm or inner membrane (IM). On the other hand, its purification to homogeneity free of IM, cell wall components, and cytoplasmic proteins is challenging. Moreover, many OMP are difficult to solubilize from the OM and often separate poorly during 2D gel electrophoresis.[5]

Presently used OM purification methods for use in mass spectrometry have involved cell breakage by French pressure cell or sonication followed by differential centrifugation and carbonate or detergent extraction of crude membranes.[1] However, for potentially lethal human pathogens like Burkholderia pseudomallei (Bp) and Burkholderia mallei (Bm) that can infect via inhalation,[6-7] sonication or other

aerosolizing cell breakage methods must be avoided. Bp causes melioidosis, a disease endemic to southeast Asia, while Bm causes glanders a disease that largely affected only horses and mules until it was eradicated from the US in the 1930s.[6-7] However, both Bm and Bp are classified as biothreat agents and so their use is highly regulated requiring BSL3-select agent containment making them difficult to work with. Thus knowledge of the constituents of their surface and OM has not been as extensively explored as many other pathogens. Nonetheless, detailed knowledge of OMPs present under various growth conditions is critical for vaccine development. Although Harding et al[7] reported identification of 35 surface proteins of Bp under one growth condition, the majority of these were predicted or documented cytoplasmic proteins; only three were predicted, likely OMPs. Moreover, most of the expected OM surface proteins such as flagellin, secretins, and TonB-type receptors were not detected in this study.

To more accurately, quantitatively, and comprehensively assess the OM proteome of Bp and Bm, we developed a safe and rapid method to highly purify OM fragments in BSL3-containment and then used trypsin shaving and LC-MS/MS to identify >170 OMP from Bm and Bp grown under a variety of conditions, including in serum and in the presence of murine macrophages. A core set of 20 OMPs that are highly abundant in the OM of Bp and Bm under all growth conditions tested was identified. It is likely these proteins are important if not essential to growth of the organism during disease and hence are prime drug or vaccine targets.

METHODS

Bacterial strains and growth conditions

B. pseudomallei strain DD503 and B. mallei ATCC 23344 were grown in the following media: 1) M9 minimal salts[8] + 3% glycerol; 2) M9 minimal salts + 3 % glycerol + 1 X BME and MEM amino acids (Sigma-Aldrich); 3.) LB[8] + 3% glycerol; 4) DMEM High Glucose (Thermo-Fisher) + 10% fetal bovine serum (GIBCO); 5) 1% glucose + 50% fetal bovine serum; 6) 3% glycerol + 3% yeast extract + 3% casamino acids; 7) DMEM High Glucose + 10% fetal bovine serum with near confluent monolayer of RAW 264.7 murine macrophages. Bacteria were grown in 250-ml flasks overnight at 37^{o} C shaking at 200 rpm, except for DMEM media where growth was unshaken with 5% CO2 in tissue culture flasks. Initial cell density was $OD_{600\ nm}$ = 0.1; cell density at harvest in to mid or late log phase was OD_{600} = 0.8-1.5, depending on media.

Outer membrane preparation

Cells were harvested by centrifugation at 7000 x g for 10 min, washed once with 0.1 vol of 20 mM Tris-HCl pH 7.0 + 3 mM $MgCl_2$ (TM) and frozen at -80°C. Pellets (200 OD_{600}) were resuspended in 3 ml of 10 mM Tris-HCl pH 7 + 25% sucrose. Lysozyme and protease inhibitor, 4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride were added to 5 mg/ml and 0.1 mg/ml, respectively. After 20 min at 37°C, $MgCl_2$ was added to 2.5 mM and incubation continued for 20 min. One vol of 4% Triton X-100 in TM was added and the tube mixed for 4 min. The lysate was frozen at -80 C and thawed at 37 ^{o} C twice with 1 min of mixing between cycles. After centrifugation at 7500 x g for 15 min, the supernatant was carefully removed and centrifuged again for 15 min. The second supernatant was removed, filtered through 0.2 μm PES syringe filter

(Whatman), and spun at 110,000 x g for 1 h in a Ti 70 rotor. The resultant supernatant was discarded and the pellet of crude outer membranes was resuspended in 0.3 ml of TM by sonicating for 5 min in a bath sonicator (Bransonic 1510). An equal volume of 4% Triton X-100 was added and the sample placed on ice for 30 min. The sample was spun at 7500 x g for 12 min and the supernatant centrifuged again at 110,000 x g for 1 h. The pelleted membranes were resuspended in 0.4 ml of TM as before and spun at 8,000 x g. The supernatant containing OM fragments (2 mg protein) was removed and frozen at -20 ^{o} C.

To remove ribosomes and loosely adhered or trapped cytoplasmic proteins, 1 mg of this partially purified OM preparation was diluted 4-fold with 5 mM Tris (pH 8), and spun at 110,000×*g* for 1 h. The pellet was resuspended by sonication in 0.6 ml 5mM Tris (pH 8) and a mixture of RNase A/RNase T1 added to 0.5 mg/ml. After incubation at 37°C for 10 min, EDTA was added to 10 mM and incubation continued for 40 min. Sample was adjusted to 0.05 M sodium carbonate and 1 M NaCl. After 1 h on ice and 15 min at 37°C, the non-vesicular membranes were pelleted at 110,000 x g for 1.25 h. The membrane pellet was washed twice by sonication in 0.1 ml of 0.1 M sodium carbonate and 1 M NaCl, incubation at 4°C for 1 hr, and centrifugation at 110,000 x g. The final yield of purified OM was 150 µg of protein as measured with a micro bicinchonic reagent method (Pierce Chemical) using 1 µl samples in 5 µl final volume. To assess purity 100 µg of membrane protein was analyzed for keto-deoxyoctanoate (KDO) using a micro-modification of the hydrolysis, periodate oxidation, and thiobartituric acid detection method of Osborn.[9-10] 100ug of membrane protein was analyzed for RNA by phenol extraction, ethanol precipitation, and gel electrophoresis.

Shotgun proteomics methods

Purified OM (50-100 µg protein) was resuspended in 0.2 ml of 50 mM ammonium bicarbonate pH 8.2 (ABC), adjusted to 10 mM dithiothreitol (DTT) and placed at 37°C for 45 min. Iodoacetamide was added to 30 mM and sample put in the dark at 25^{o} C for 30 min after which 0.2 ml of 50 mM ABC was added. The reduced alkylated membranes were spun at 110,000 x g for 1.25 h. The pellet was suspended by sonication in 0.2 ml of 50 ABC, treated with 5 µg of modified trypsin (Promega) at 37°C overnight and then centrifuged 110,000 x g for 1.25 h. The pellet was sonicated 5 min in 0.2 ml of 50 mM ABC, placed at 90°C for 20 min, then quick chilled on ice. Methanol (300 µl) was added followed by 5 µg of trypsin and incubation overnight at 37°C. The sample was centrifuged at 110,000 x g for 1.25 h and the supernatant containing 15 µg of peptides was removed, vaccum dried, dissolved in 0.1% trifluoroacetic acid (TFA), and centrifuged at 10,000 x g for 5 min. The supernatant was vacuum dried, dissolved in 0.1% formic acid in 2% acetonitrile (ACN), and loaded on a 75 µm x 105 mm C18 reverse phase column (packed in house, YMC GEL ODS-AQ120ÅS-5, Waters) by nitrogen bomb. Peptides were eluted directly into the nanospray source of a linear ion trap (Thermo Finnigan LTQ™)[11] with a 160-min linear gradient of 4 to 80% ACN in 0.1% formic acid over 100 min at a flow rate of ~250 nl/min. Full scan MS spectra were acquired from m/z 300 to 2000 followed by 8 MS/MS events of the most intense ions. A dynamic exclusion window was applied which prevents the same m/z value from being selected for 12 seconds after its acquisition. Data were automatically acquired using Xcalibur® (ver. 2.0.7, Thermo Fisher Scientific), subsequently analyzed using SEQUEST (Bioworks 3.3, Thermo Fisher Scientific)[12], and finally filtered to achieve a

1% false discovery rate on the protein level using ProteoIQ™[13]. The abundance of each identified protein was estimated from the normalized spectral counts, which are calculated as the number of spectral counts for each protein (SpC) divided by its number of amino acids (L) divided by the sum of SpC/L for all proteins in the experimental dataset.

Peptide mass fingerprinting

Gel slices were incubated in 0.3 ml of 50 mM ammonium bicarbonate (ABC)/50% methanol at 37°C for 30 min and supernatant discarded. This incubation/washing was repeated twice and then 0.5 ml 100% ACN was added. After 15 min ACN was discarded, gel slices dried in a spinvac and then incubated in 10 mM DTT in 20 mM ABC at 60°C for 20 min, followed by 100 mM iodoacetamide in 20 mM ABC at 25^{o} C for 30 min. This was followed by treatment twice with 50 mM ABC/50% methanol at 25^{o} C for 15 min. After the supernatant was removed, gel slices were dried in a speedvac and treated with 1 μg of trypsin in 50 μl of 20 mM ABC/10% ACN for 18 h at 37°C. The supernatant was removed and gel slices were incubated with 60 μl of 50% ACN/0.1% TFA at 25^{o} C for 40 min. Both peptide containing supernatants were combined, dried in the spinvac to 10 μl and then 35 μl of 0.1% TFA added. A Glygen NuTip C-18 tip was activated with 100% CAN, washed with 0.1% TFA, and peptide samples bound to the tips by filling and expulsion 20 times. The tips were washed with 0.1% TFA and the peptides eluted with 2.5 μl of MALDI matrix onto a MALDI plate and run on a Bruker Daltronics Autoflex mass spectrometer.[14]

Bioinformatic methods

Using keywords such as outer membrane, lipoprotein, flagella, pili, secretin, porin, receptor, and surface at IMG (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) we manually assembled a list of COG numbers containing proteins associated with the OM. Using this list we extracted a list of 150 predicted OMPs of Bp. Similarly we extracted a list of all 140 Bp proteins scoring >7 in the outer membrane localization category in PsortB (http://www.psort.org/genomes/genomes.pl). Approx 112 proteins were found in common on both lists yielding a list of highly probable OMPs of Bp. Lists were combined and redundant entries removed resulting in a list of 175 potential membrane proteins.

RESULTS AND DISCUSSION

Purification and characterization of OM fragments

To isolate an OM-enriched fraction, Bp or Bm cells were treated with lysozyme and then freeze-thawed in the presence of Triton X-100 and $MgCl_2$. This lysed the cells and solubilized proteins of the cytoplasm, periplasm, and IM. The insoluble OM fragments were recovered from this lysate by differential centrifugation. Initial LC-MS/MS analyses of tryptic peptides from these crude OM preparations identified > 300 proteins. However, a significant portion of these were abundant, documented cytoplasmic proteins (e.g. enzymes of glycolysis and TCA cycle, amino acyl-tRNA synthetases, RNA polymerase subunits, etc.). IM proteins were in relatively low abundance, but many ribosomal proteins and RNA were detected suggesting ribosomal contamination. To remove any ribosomes and proteins adsorbed to or trapped inside vesicular membranes, the OM-enriched fraction was treated with RNase and EDTA, and

then subjected to extensive washing with 1 M NaCl and 0.1 M sodium carbonate pH 11. This resulted in a dramatic reduction in the apparent contamination with cytoplasmic proteins and removal of all rRNA. In contrast to the crude OM preparations that showed keto-deoxyoctanoate (KDO) to protein ratios of 0.01 umol/mg protein, ratios for this RNase-NaCl-pH 11 treated OM fraction was 0.03, near the value of 0.04 reported for highly purified OM of Bp.[15] Furthermore SDS-PAGE analysis (Figure C-1) showed the purified OM preparation contained 5 major protein species that comprised the vast majority of Coomassie Blue stained material. Banding pattern is similar to that of the highly purified Bp OM preparation obtained by Gotoh et al. using sucrose gradient purification.[15] Peptides from each band were obtained by trypsin treatment and analyzed by MALDI-TOF MS. Four of the five bands were identified as containing probable OMPs: BPSL0294, BPSL0894, and BPSL2522 that are predicted porins and BPSL1775, a predicted iron uptake receptor. Taken together these data indicated that the OM preparation was highly enriched for OM proteins of Bp. The fuzzy diffuse fifth band with a high molecular mass (150 kDa; Figure C-1) produced no peptides after trypsin digestion consistent with a polysaccharide composition.

Preliminary LC-MS/MS analyses

The highly purified OM fraction was incubated with trypsin and the released (shaved) peptides analyzed by LC-MS/MS. Over 225 proteins were identified at >99% confidence. Several of the most abundant proteins identified were the same OM proteins detected on the SDS-PAGE gel; however, peptides from some ribosomal and other abundant cytoplasmic proteins were still detected at moderate to low levels. Inner membrane associated proteins such as NADH dehydrogenase, F1-F0 ATPase,

cytochrome oxidases, or permeases[16] were conspicuously absent. These surface-shaved membranes which should be reduced in non-specifically adsorbed proteins were then washed twice with 1M NaCl at pH 11, dissolved in 60% methanol and re-treated with trypsin to release peptides from OMPs that were more deeply imbedded in the lipid bilayer. LC-MS/MS analyses of these peptides identified ~200 proteins. The majority appeared to be likely OM proteins, but 25% were abundant well-known cytoplasmic proteins (e.g. some ribosomal proteins, subunits of RNA polymerase, pyruvate dehydrogenase, oxoglutarate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, translation elongation factors, amino acyl-tRNA synthetases, glutamine/glutamate synthetases). Others have reported that ribosomal and some other highly abundant proteins like these strongly associate with the bacterial membrane surface[17-19] and some have suggested this to be biologically relevant[19]. More likely these highly abundant cytoplasmic proteins (or peptides derived from them) strongly adhere to the OM via hydrophobic or electrostatic interaction as an artifact of preparation and were not included in later analyses.

Elucidation and Analysis of OM proteome of B. pseudomallei

We purified and analyzed OM fractions as described above from Bp cells grown under 7 different growth/media conditions. Proteins identified in OM preparation were pooled, duplicates removed, and a non-redundant list of 160 OMPs detected in any of the 7 preparation assembled and designated the OM proteome of Bp. Of these 45% were predicted by PsortB to be OMPs and 61 were found in the comprehensive list of 175 in silico predicted OMPs of Bp. Nearly half of the 115 OMPs predicted by both IMG and PsortB were detected. This proportion is similar to what was reported by Chung et al.[2] for

Actinobacillus pleuropnemoniae and Babujee et al. for Erwinia chrysanthemi[1]. Similar results were obtained from purified OM of Bm (see below).

Of the 31 porins that comprise the most populous group of predicted and observed OMPs of Bp (COG3202), we detected 13, nearly twice that of any previous study.[1-2, 18] Porins form water-filled channels that allow the diffusion of hydrophilic molecules across the OM into the periplasm; some allow any molecule that fits to pass through the channel, while others are specific via binding sites inside the pores (e.g. BPSL0294 carbohydrate selective porin). Another populous family of OMPs is OM efflux channels (COG 1538/PFAM 2321). These proteins form trimeric channels with a 12 stranded beta barrel that spans the OM and is coupled to a helical barrel spanning the periplasm to allow export of diverse substrates[20]; we detected 11 of the 21 efflux channels predicted in Bp. Members the OmpW and OmpA families are also predicted to be prominent in the OM playing multifunctional roles which are mechanistically unclear; we detected 4 of 9 predicted OmpAs and 3 of 5 predicted OmpWs. The expected structural components of the flagella and its basal body: flagellin (BPSL3319), L and P ring proteins (BPSL0276 & BPSL0277), hook and rod proteins BPSL0273 and BPSL02750 were also detected.

Metal acquisition, especially of iron, is important both for saprophytic and pathogenic bacteria. It often involves OM-surface receptors that bind a chelated metal complex (e.g. Fe-siderophores). The complex is then imported across the OM using TonB and a membrane coupling protein (MCP). Receptor proteins of this type are assigned to COGs 1629, 4773, and 4774; all have PFAM domain 0593. Bp has 8 proteins assigned to these COGs of which we detected all 8, including the ones predicted to import heme, copper, iron, and a pyochelin-like Fe-siderophore. One of these, BPSL1775, encoding a

predicted catechol-based Fe-siderophore receptor, was the fifth most abundant protein in the OM and was detected in cells from all growth conditions (see below); not unexpectedly it reacts with convalescent sera from Bp-infected patients.[21-22]

Extracellular polysaccharides (EPS) are important virulence factors that protect Bm, Bp and other pathogens from host recognition.[23-25] Bp has 5 predicted proteins that are similar to Wza the probable translocon channel for secretion of colanic acid EPS through the OM of E. coli.[26] Although Bp has been reported to produce only 3 types of EPS[27] this suggests it may have the capacity to produce 5 distinct EPSs. Three of these 'EPS translocon channel proteins' in COG1596 were detected in the Bp OM: BPSS0417 in the lipopolysaccharide O-antigen cluster,[23] BPSL2807 in the wbc capsule-associated gene cluster[25] and BPSS1831 in a cluster of predicted polysaccharide biosynthesis genes of unknown function.

The OM is not only the primary barrier against host defenses but is also site of many offensive weapons Bp uses against the host, including secretin pores of Type II and Type III secretion systems that function in virulence and pathogenicity of Bp.[28-30] Its genome encodes 4 secretin-like proteins in COG 1450, each predicted to be a pore/channel for a different type secretion system involved in export of proteins across the OM. We detected 3 of these in our OM preparation, one (BPSS1545, BsaO) from in the Type III secretion system required for virulence[29] another BPSL0007 (GspD) is encoded in the gene cluster for the general (Type II) secretion pathway involved in virulence[28] and the third, BPSS1600, is likely involved in Type IV pilus assembly which functions in twitching motility. BPSL0007 was among the 20 most abundant proteins detected in OM preparations from cells grown in all conditions (see below). Interestingly,

BPSL0007 and BPSL2807, one of the EPS translocon channel proteins discussed above strongly reacted with sera from recovered melioidosis patients indicating these are expressed in vivo.[21-22]

We detected multiple components of a Type VI secretion system (T6SS).[31-33] Nearly all these derived from only one of the six distinct T6SS gene clusters in Bp. They were not from the T6SS cluster required for virulence but rather from the T6SS gene clusters (systems) that is orthologous to T6SSs clusters present in several dozen pathogenic and nonpathogenic proteobacteria. Two of the Bp T6SS components, BPSL3105 and BPSL3107 were among the top 15 most abundant OMPs detected under all growth conditions (below). BPSL3105 encodes the so-called HCP protein that forms a ring-like structure and was proposed to assemble into a hollow secretion tube that passes from the periplasm through the OM to the cell surface.[32-34] The other abundant T6SS component detected, BPSL3107 (COG3516), has not been ascribed a functional role but has been proposed to interact with the IcmF-like protein of the T6SS (BPSL3097) which we also found in the OM. BPSL3097 has been proposed to interact with a lipoprotein like BPSS3108 also found in the BP OM. We also detected BPSS0078/BPSS1213), a VgrG-like needle protein and BPSL3101, a ClpV-related ATPase, in the OM; VgrG has been proposed to interact with the end of the HCP tube and assist in penetration of the host cell while the ClpV-related ATPase has been proposed to unfold T6SS substrates and provide the energy for their export across the OM.[34] These observations support the structural model of the Type VI secretion machine proposed by Leiman et al.[35] However, we also found other proteins from the same T6SS gene cluster in the OM which we not part of this model: BPSL3106-BPSL3110 and

BPSL3103. So it will be necessary to adapt the model to include these 'new' parts. Although the detailed structure and topology of the T6SS secretion machine remains to be fully elucidated, our results suggest that proteins in COGs 3501 (Rhs/Vgr family), 3516 (ImpB/EvpA family), 3157 (HCP family), 3517 (Evp/ImpD family), 3519 (ImpG family), 3522 (ImpJ family), 3523 (IcmF/VasK family), and maybe 5295, comprise its major OM components; some of these undoubtedly are surface exposed. Consistent with this BPSS0078 (COG3501) and BPSL3103 (COG3519) react with convalescent sera of Bp-infected patients.[21-22] We did not detect any predicted T6SS proteins from COG3515 (ImpA family), COG3520, COG3521 (FHA family) COG3456, COG3455 (DotU family) or COG3518; these maybe loosely attached components of the T6SS apparatus or involved only in assembly of the T6SS machine.

We identified 14 'hypothetical' proteins in the Bp OM, nine of which were assigned to COGs containing proteins associated with OM functions. The other 5 had no COG assignment or PFAM domain, but 3 were predicted by PSORTB to be OMPs. Two of these, BPSS1996 and BPSL2151, are expressed in vivo (i.e react with sera from Bp-infected patients).[21-22] BPSL2151 harbors a PFAM domain found in the OMP85 family of proteins that facilitate insertion of lipids and proteins into the OM (ref). These 'hypothetical' proteins likely represent examples of new families of OMPs with unique functions that remain to be elucidated.

Identification of proteins always present in the OM of Bp

From the lists of OMPs detected under each of the 7 growth conditions used (see Methods), we derived a list of OMPs detected in every preparation irrespective of growth conditions. The 20 most abundant of these 'always present' OMPs are listed in Table C-1

in order of abundance in the OM of cells grown in defined minimal medium with glycerol as the sole carbon source and only added organic molecule. As estimated from spectral counts, the amount of these 20 OMPs in the Bp OM is >50% of total OMPs present. The fact that these are highly expressed and abundantly present in cells growing in such a wide variety of conditions (e.g. in minimal and rich medium, in serum, in presence of macrophages) implies they are fundamental and essential to Burkholderia physiology in many, if not all, environments. Sixteen of these 20 proteins were also detected in preparations of the Bm OM (see below). The majority are porins and OmpW-like proteins. These common and abundant OMPs should be good targets for multivalent vaccines or drug development, in particular porins BPSS0879 and BPSS167 that react with convalescent sera from Bp-infected patients[21-22]. Another always present OMP is the Type II secretion pore GspD (BPSL0007) and BPSL3105 the tube-forming component (Hcp) of a T6SS. Two of the top 7, BPSS1356 and BPSL2003, are hypothetical proteins. The former is a 120-kDa protein found by BLAST to only have orthologs in other Burkholderia sp. The latter is a 12-kDa protein in COG0393 found by BLAST to be present in all sequenced Burkholderia strains and a few close relatives. No PFAM domains are evident in these novel abundant OMPs so their role in the OM remains completely unknown. Interestingly, when purified amino acids were added to the growth media the relative abundance of 2 porins, BPSL0289 and BPSL0294, increased > 6-fold, suggesting they may be involved in amino acid uptake.

Elucidation of the OM proteome of B. mallei

B. mallei (Bm) is a deletion clone of Bp which has lost >1200 genes by Is-mediated deletion.[36] With 3 exceptions, all genes of B. mallei are found in Bp with

>99.7% DNA sequence identity.[37] The massive gene lost makes Bm less metabolically/physiologically diverse, robust and capable than Bp. We analyzed the Bm OM proteome as we did for Bp; however, because they grew so poorly in minimal media without amino acids, in serum, and in the presence of macrophages we could not analyze Bm OM preparations from these conditions. The nonredundant list of proteins detected in the Bm OM from cells grown in any of the 4 conditions where they grew almost as well as Bp, contained 105 proteins. Orthologs of 72 of these were also detected in the Bp OM; these 72 likely represent the most fundamental OMPs of these two pathogens. We also generated a list of the 20 most abundant OMPs detected in Bm cells regardless of growth condition (Table C-2). Orthologs of 6 were present in the analogous list of abundant Bp OMPs (Table C-1). Orthologs of the remaining 14 were detected in the Bp OM, but in lower abundance and sometimes not from all growth conditions. While 9 of the most abundant OMPs of Bp were porins, only 4 porins are found in the 20 most abundant OMPs of Bm. Moreover, 3 of the most abundant Bm OMPs were TonB-dependent receptor proteins involved in metal, in contrast to only one in the Bp list. The disparity between amounts and identities of the 20 most abundant omnipresent OMPs of Bp and Bm can be partially explained by the fact that Bp produces OMPs whose genes have been deleted from the Bm genome, thus skewing the relative abundance determinations. For example, the genes encoding the orthologs of the T6SS components that were very abundant in the Bp OM are missing from the Bm genome. Another likely reason for the disparity is that because Bm has become a host-adapted pathogen and is no longer capable of environmental survival like Bp, it expresses a somewhat different spectrum of OMPs more appropriate for survival in animal hosts. For example, the BMAA0749 and

BMA0729.1 proteins of the T6SS required for Bm virulence in hamsters[31] were detected in Bm OM preparations, but not in those of Bp. Several other proteins found in the Bm OM, but not in the Bp OM, are predicted surface proteins possibly related to virulence: BMAA0251, a TonB-dependent receptor, and BMAA1936 and BMA1701 which are lipoproteins related to a Rickettsial surface antigen. Another is BMAA2738 which could be involved in coping with oxidative stress.

CONCLUSION

A Triton-insoluble fraction of lysozyme treated cells of Bp and Bm was treated with RNase and repeatedly extracted with high salt at pH 11 yielding a highly purified OM preparation as evidenced by its KDO to protein ratio and SDS-PAGE analysis. This preparation was as pure as that of Gotoh et. al[15] who used a more hazardous and laborious method involving sonication and sucrose gradient centrifugation. Trypsin shaving of the OM from cells grown under various conditions followed by LC-MS/MS identified >190 possible OMPs. This is significantly more than the 30 to 60 OMPs identified in most previous OM proteome determinations of proteobacteria that relied on 2D gel fractionation of OMs purified away from proteins of the IM and cytoplasm by either a single detergent or pH 11 extraction and MALDI-TOF[1-4]. Of the 160 proteins we identified in the Bp OM, 100 were predicted by PSORTB or IMG to be likely OMPs. Of the remaining 60 about half were proteins whose predicted functions are associated with cell envelope-related functions: cell division, cell wall biosynthesis, flagella, and transport; some of these may represent new classes or families of OMPs, but further experimentation is required for confirmation.

Only one other OM proteome determination reported more OMP identifications than here; however, it employed a much more laborious combination of 1-D and 2-D gel separation followed by LC-MALDI-TOF-TOF analysis of dozens gel slices.[38] Because we directly analyzed total peptides shaved from highly purified OM preparations in an LC-MS/MS ion trap instrument without prior electrophoresis and inefficient peptide extraction from gel pieces, we were more able to analyze multiple samples from cells grown under a wide variety of conditions, with a higher sensitivity and also estimate relative abundances. This allowed us to assemble a list of 20 'omnipresent' OMPs that are present and abundant in Bm and Bp under multiple culturing regimes. Many of these proteins are likely to be indispensible for growth/survival and accessible on the surface of Bp and Bm cells in many environments, and thus should be ideal targets for therapeutics and vaccines. In support of this 20% of the identified OMPs and 2 of the 5 most abundant OMPs have been found to react with sera of recovered melioidosis patients indicating they are expressed in vivo during pathogenesis.[21-22] However, it should be noted that nearly half the predicted OMPs of Bp were not detected in any preparation; this was particularly obvious for porins, efflux pumps, and secretins. When and where these are expressed remains to be determined.

Finally, we made several novel observations about the probable membrane topology of 10 of the 15 components of the recently discovered T6SSs that in some cases are required for pathogenesis. For the most part, these findings experimentally confirmed predictions and models based on structural analyses of T6SS proteins.[34] However, several T6SS-associated proteins not accounted for in previous models were detected and hence require further study and integration in the models for T6SS structure and functions.

REFERENCES

1. Babujee, L.; Venkatesh, B.; Yamazaki, A.; Tsuyumu, S., Proteomic analysis of the carbonate insoluble outer membrane fraction of the soft-rot pathogen Dickeya dadantii (syn. Erwinia chrysanthemi) strain 3937. *J Proteome Res* **2007,** 6, (1), 62-9.

2. Chung, J. W.; Ng-Thow-Hing, C.; Budman, L. I.; Gibbs, B. F.; Nash, J. H.; Jacques, M.; Coulton, J. W., Outer membrane proteome of Actinobacillus pleuropneumoniae: LC-MS/MS analyses validate in silico predictions. *Proteomics* **2007,** 7, (11), 1854-65.

3. Boyce, J. D.; Cullen, P. A.; Nguyen, V.; Wilkie, I.; Adler, B., Analysis of the Pasteurella multocida outer membrane sub-proteome and its response to the in vivo environment of the natural host. *Proteomics* **2006,** 6, (3), 870-80.

4. Berven, F. S.; Karlsen, O. A.; Straume, A. H.; Flikka, K.; Murrell, J. C.; Fjellbirkeland, A.; Lillehaug, J. R.; Eidhammer, I.; Jensen, H. B., Analysing the outer membrane subproteome of Methylococcus capsulatus (Bath) using proteomics and novel biocomputing tools. *Arch Microbiol* **2006,** 184, (6), 362-77.

5. Cordwell, S. J., Technologies for bacterial surface proteomics. *Curr Opin Microbiol* **2006,** 9, (3), 320-9.

6. Currie, B. J., Burkholderia pseudomallei and Burkholderia mallei: Melioidosis and Glanders. In *Principles and Practice of Infectious Diseases*, 2 ed.; Mandell, G. L., Bennett, J.E., Dolin, R., Ed. Churchill Livingstone: New York, 2005; pp 2622-2632.

7. Harding, S. V.; Sarkar-Tyson, M.; Smither, S. J.; Atkins, T. P.; Oyston, P. C.; Brown, K. A.; Liu, Y.; Wait, R.; Titball, R. W., The identification of surface proteins of Burkholderia pseudomallei. *Vaccine* **2007,** 25, (14), 2664-72.

8. Miller, J., *Methods Experiments in Molecular Genetics*. 1972; p 431-433.

9. Osborn, M. J., Studies on the Gram-Negative Cell Wall. I. Evidence for the Role of 2-Keto- 3-Deoxyoctonate in the Lipopolysaccharide of Salmonella Typhimurium. *Proc Natl Acad Sci U S A* **1963,** 50, 499-506.

10. Weissbach, A.; Hurwitz, J., The formation of 2-keto-3-deoxyheptonic acid in extracts of Escherichia coli B. I. Identification. *J Biol Chem* **1959,** 234, (4), 705-9.

11. McLuckey, S. A.; Van Berkel, G. J.; Goeringer, D. E.; Glish, G. L., Ion trap mass spectrometry. Using high-pressure ionization. *Anal Chem* **1994,** 66, (14), 737A-743A.

12. Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D., Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **1995,** 67, (8), 1426-36.

13. Weatherly, D. B.; Atwood, J. A., 3rd; Minning, T. A.; Cavola, C.; Tarleton, R. L.; Orlando, R., A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics* **2005,** 4, (6), 762-72.

14. Hillenkamp, F.; Karas, M.; Beavis, R. C.; Chait, B. T., Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem* **1991,** 63, (24), 1193A-1203A.

15. Gotoh, N.; White, N. J.; Chaowagul, W.; Woods, D. E., Isolation and characterization of the outer-membrane proteins of Burkholderia (Pseudomonas) pseudomallei. *Microbiology* **1994,** 140 (Pt 4), 797-805.

16. Kadner, R., The cytoplasmic membrane. In *Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology*.

17. Tjalsma, H.; Lambooy, L.; Hermans, P. W.; Swinkels, D. W., Shedding & shaving: disclosure of proteomic expressions on a bacterial face. *Proteomics* **2008,** 8, (7), 1415-28.

18. Molloy, M. P.; Phadke, N. D.; Maddock, J. R.; Andrews, P. C., Two-dimensional electrophoresis and peptide mass fingerprinting of bacterial outer membrane proteins. *Electrophoresis* **2001,** 22, (9), 1686-96.

19. Wei, C.; Yang, J.; Zhu, J.; Zhang, X.; Leng, W.; Wang, J.; Xue, Y.; Sun, L.; Li, W.; Jin, Q., Comprehensive proteomic analysis of Shigella flexneri 2a membrane proteins. *J Proteome Res* **2006,** 5, (8), 1860-5.

20. Buchanan, S. K., Type I secretion and multidrug efflux: transport through the TolC channel-tunnel. *Trends Biochem Sci* **2001,** 26, (1), 3-6.

21. Su, Y. C.; Wan, K. L.; Mohamed, R.; Nathan, S., A genome level survey of Burkholderia pseudomallei immunome expressed during human infection. *Microbes Infect* **2008,** 10, (12-13), 1335-45.

22. Felgner, P. L.; Kayala, M. A.; Vigil, A.; Burk, C.; Nakajima-Sasaki, R.; Pablo, J.; Molina, D. M.; Hirst, S.; Chew, J. S.; Wang, D.; Tan, G.; Duffield, M.; Yang, R.; Neel, J.; Chantratita, N.; Bancroft, G.; Lertmemongkolchai, G.; Davies, D. H.; Baldi, P.; Peacock, S.; Titball, R. W., A Burkholderia pseudomallei protein microarray reveals

serodiagnostic and cross-reactive antigens. *Proc Natl Acad Sci U S A* **2009,** 106, (32), 13499-504.

23. DeShazer, D.; Brett, P. J.; Woods, D. E., The type II O-antigenic polysaccharide moiety of Burkholderia pseudomallei lipopolysaccharide is required for serum resistance and virulence. *Mol Microbiol* **1998,** 30, (5), 1081-100.

24. DeShazer, D.; Waag, D. M.; Fritz, D. L.; Woods, D. E., Identification of a Burkholderia mallei polysaccharide gene cluster by subtractive hybridization and demonstration that the encoded capsule is an essential virulence determinant. *Microb Pathog* **2001,** 30, (5), 253-69.

25. Reckseidler, S. L.; DeShazer, D.; Sokol, P. A.; Woods, D. E., Detection of bacterial virulence genes by subtractive hybridization: identification of capsular polysaccharide of Burkholderia pseudomallei as a major virulence determinant. *Infect Immun* **2001,** 69, (1), 34-44.

26. Dong, C.; Beis, K.; Nesper, J.; Brunkan-Lamontagne, A. L.; Clarke, B. R.; Whitfield, C.; Naismith, J. H., Wza the translocon for E. coli capsular polysaccharides defines a new class of membrane protein. *Nature* **2006,** 444, (7116), 226-9.

27. Kawahara, K.; Dejsirilert, S.; Ezaki, T., Characterization of three capsular polysacharides produced by Burkholderia pseudomallei. *FEMS Microbiol Lett* **1998,** 169, (2), 283-7.

28. DeShazer, D.; Brett, P. J.; Burtnick, M. N.; Woods, D. E., Molecular characterization of genetic loci required for secretion of exoproducts in Burkholderia pseudomallei. *J Bacteriol* **1999,** 181, (15), 4661-4.

29. Warawa, J.; Woods, D. E., Type III secretion system cluster 3 is required for maximal virulence of Burkholderia pseudomallei in a hamster infection model. *FEMS Microbiol Lett* **2005,** 242, (1), 101-8.

30. Ulrich, R. L.; DeShazer, D., Type III secretion: a virulence factor delivery system essential for the pathogenicity of Burkholderia mallei. *Infect Immun* **2004,** 72, (2), 1150-4.

31. Schell, M. A.; Ulrich, R. L.; Ribot, W. J.; Brueggemann, E. E.; Hines, H. B.; Chen, D.; Lipscomb, L.; Kim, H. S.; Mrazek, J.; Nierman, W. C.; Deshazer, D., Type VI secretion is a major virulence determinant in Burkholderia mallei. *Mol Microbiol* **2007,** 64, (6), 1466-85.

32. Mougous, J. D.; Cuff, M. E.; Raunser, S.; Shen, A.; Zhou, M.; Gifford, C. A.; Goodman, A. L.; Joachimiak, G.; Ordonez, C. L.; Lory, S.; Walz, T.; Joachimiak, A.; Mekalanos, J. J., A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. *Science* **2006,** 312, (5779), 1526-30.

33. Pukatzki, S.; Ma, A. T.; Sturtevant, D.; Krastins, B.; Sarracino, D.; Nelson, W. C.; Heidelberg, J. F.; Mekalanos, J. J., Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. *Proc Natl Acad Sci U S A* **2006,** 103, (5), 1528-33.

34. Pukatzki, S.; Ma, A. T.; Revel, A. T.; Sturtevant, D.; Mekalanos, J. J., Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. *Proc Natl Acad Sci U S A* **2007,** 104, (39), 15508-13.

35. Leiman, P. G.; Basler, M.; Ramagopal, U. A.; Bonanno, J. B.; Sauder, J. M.; Pukatzki, S.; Burley, S. K.; Almo, S. C.; Mekalanos, J. J., Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. *Proc Natl Acad Sci U S A* **2009,** 106, (11), 4154-9.

36. Nierman, W. C.; DeShazer, D.; Kim, H. S.; Tettelin, H.; Nelson, K. E.; Feldblyum, T.; Ulrich, R. L.; Ronning, C. M.; Brinkac, L. M.; Daugherty, S. C.; Davidsen, T. D.; Deboy, R. T.; Dimitrov, G.; Dodson, R. J.; Durkin, A. S.; Gwinn, M. L.; Haft, D. H.; Khouri, H.; Kolonay, J. F.; Madupu, R.; Mohammoud, Y.; Nelson, W. C.; Radune, D.; Romero, C. M.; Sarria, S.; Selengut, J.; Shamblin, C.; Sullivan, S. A.; White, O.; Yu, Y.; Zafar, N.; Zhou, L.; Fraser, C. M., Structural flexibility in the Burkholderia mallei genome. *Proc Natl Acad Sci U S A* **2004,** 101, (39), 14246-51.

37. Losada, L.; Ronning, C. M.; DeShazer, D.; Woods, D.; Fedorova, N.; Kim, H. S.; Shabalina, S. A.; Pearson, T. R.; Brinkac, L.; Tan, P.; Nandi, T.; Crabtree, J.; Badger, J.; Beckstrom-Sternberg, S.; Saqib, M.; Schutzer, S. E.; Keim, P.; Nierman, W. C., Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. *Genome Biol Evol* **2010,** 2, 102-16.

38. Veith, P. D.; O'Brien-Simpson, N. M.; Tan, Y.; Djatmiko, D. C.; Dashper, S. G.; Reynolds, E. C., Outer membrane proteome and antigens of Tannerella forsythia. *J Proteome Res* **2009,** 8, (9), 4279-92.

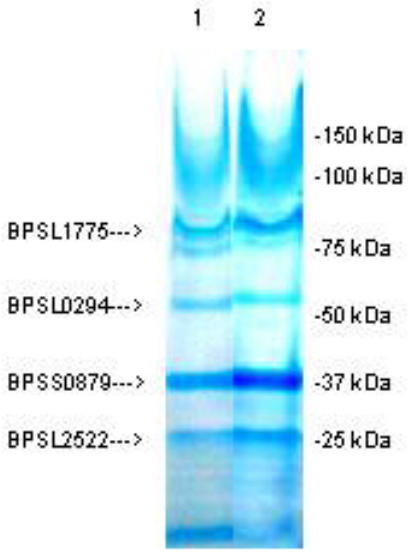


Figure C-1. SDS-PAGE and peptide mass fingerprint analysis of a purified OM preparation from B. pseudomallei.
10 µg (lane 1) and 40 µg (lane 2) of protein from a Triton insoluble OM pellet that had been further purified by extraction with RNase-NaCl-pH 11 was boiled for 5 min in sample denaturing buffer (10 mM Tris-HCl pH 6.8, 2% SDS, 2% mercaptoethanol) and electrophoresed at 100 V for 3 hrs on a 4 to 20% gradient polyacrylamide gel (ThermoFisher 25204) before staining with Coomassie Blue R250 (Biorad). Bands were excised and the protein species present in them identified by peptide mass fingerprinting as outlined in Methods. Bp cells were grown in M9 minimal salts media with glucose and amino acids.

Table C-1. 20 most abundant OM proteins of B. pseudomallei detected under all growth conditions.

Bp Locus Tag[f]	Locus Tag of BM Ortholog	Predicted Function of Gene Product	COG Assignment	% of Total OMPs[b]	Found in Bm OM[c]	% of Total OMPs[d]
BPSL2559[a]	BMA2089	porin protein	3203	18.78	Y	6.82
BPSL2989	BMA2507	lipoprotein	3133	12.96	Y	5.23
BPSL2704	BMA2010	OmpW-family protein	3407	5.88	Y	0.37
BPSS0879[a,e]	BMAA1353	porin protein	3203	4.63	Y	11.09
BPSS1356	BMAA0912	hypothetical protein	3246	3.69	Y	1.10
BPSL2522	BMA0436	porin protein in OmpA family	2885	3.53	Y	7.63
BPSL2003	BMA0904	hypothetical protein	393	1.78	Y	1.05
BPSS1679	BMAA1698	porin protein	3203	1.67	Y	1.47
BPSS0943	BMAA1286	porin protein	3203	1.42	Y	3.10
BPSL1029	BMA0743	porin protein	3203	1.08	Y	1.47
BPSL1775[a,e]	BMA1178	putative Fe uptake receptor 'for monomeric catechols'	4774	0.79	Y	3.89
BPSL3036	BMA2562	porin protein	3203	0.77	Y	1.06
BPSL3161	BMA2738	OmpW-family protein	3047	0.61	N	0.50
BPSL0816	BMA0317	oprB OM efflux protein	1538	0.58	Y	0.48
BPSL3105	None	Type VI secretion associated HCP protein	3157	0.30	n/a	0.48
BPSL3107	None	Type VI secretion associated protein	3516	0.29	n/a	0.22
BPSL0289	BMA3345	porin protein	3203	0.26	Y	1.87
BPSL0007[e]	BMA2786	gspD general secretory pathway secretin protein	1450	0.20	Y	0.22
BPSL0294[a]	BMA3354	protein in carbohydrate selective porin family	3659	0.19	Y	1.25
BPSL1913	BMA1056	OM lipoprotein in NodT family	1538	0.20	N	0.63
		TOTAL		60.3		49.7

a---detected on SDS-PAGE gel (Fig. 1)

b---Cells grown in M9 minimal media with 3% glycerol; relative abundance determined by spectral counts

c---Y, yes; N, no; n/a, no ortholog in Bm genome

d---Cells grown in M9 salts media with 3% glycerol plus 20 amino acids; abundance determined by spectral counts

e---reported to reacted with sera from Bp infected patients

f--- from Integrated Microbial Genomes

Table C-2. 20 most abundant OM proteins of B. mallei detected under all growth conditions.

Bm Locus Tag	Locus Tag of Bp Ortholog	Predicted Function of Gene Product	COG Assignment	% of Total OMPs[a]
BMA3354	BPSL0294	carbohydrate porin, OprB family	3659	15.28
BMA0436	BPSL2522	OmpA family porin protein	2885	10.59
BMA1178	BPSL1775	TonB-dependent siderophore receptor	4774	5.22
BMAA1826	BPSS0244	TonB-dependent heme/hemoglobin receptor family protein	1629	4.29
BMAA1353	BPSS0879	porin OpcP	3203	2.46
BMA0317	BPSL0816	RND efflux system, OM lipoprotein, NodT family	1538	2.15
BMA0465	BPSL2543	lipoprotein	None	1.49
BMAA1464	BPSS0294	RND efflux system, OM lipoprotein, NodT family	1538	1.33
BMA2307	BPSL2807	capsular polysaccharide biosynthesis/export protein	1596	1.19
BMA1547	BPSL2151	OM protein, OMP85 family	4775	1.07
BMA0208	BPSL0658	organic solvent tolerance protein	1452	0.52
BMAA0356	BPSS2136	porin, OprD family	None	0.51
BMA0685	BPSL0976	vitamin B12 receptor BtuB	4206	0.50
BMAA2092	BPSS2331	lipoprotein	2853	0.47
BMA2723	BPSL3147	lipoprotein	2853	0.33
BMA2786	BPSL0007	general secretion pathway protein D	1450	0.28
BMAA0427	BPSS1742	TonB-dependent copper receptor	1629	0.22
BMA0316	BPSL0815	hydrophobe/amphiphile efflux family protein	841	0.14
BMA0705	BPSL0994	OM protein, OMP85 family	729	0.07
BMAA0486	BPSS0562	porin protein	3203	0.03
		TOTAL		48.2

a---cells were grown in M9 minimal media with glycerol and amino acids; relative abundance among total OMPs determined by spectral counts