THE APPLICATION OF MASS SPECTROMETRY-BASED LABEL-FREE QUANTITATIVE PROTEOMIC STRATEGIES IN CANCER STEM CELL RESEARCH

by

Lan Dai

A dissertation submitted in partial fulfillment
Of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
In The University of Michigan
2011

Doctoral Committee:

Professor David M. Lubman, Chair
Professor Philip C. Andrews
Associate Professor Kerby A. Shedden
Associate Professor Kristina I. Hakansson
Assistant Professor Subramaniam Pennathur

To my family

ACKNOWLEDGEMENTS

Life is a journey. My six and a half years Ph.D training period has taught me how to cruise through from dusk to dawn.

I must first thank my advisor professor David M. Lubman for providing me everything necessary for my graduate school studies. His passion for research and his deep insights into biological mass spectrometry wowed me countless times. I would also like to thank Professor Shedden, Professor Pennathur, Professor Andrews, and Professor Hakansson for serving on my committee and for giving me valuable advice on my research. I also owe my gratitude to the whole lab where I have spent 6 years. People come and go, but the spirits and the wisdom are retained and transferred.

Finally, I would like to give my special thanks to my husband and my mother. They have magically managed to put up with my sometimes ‘irrational’ behavior during ups and downs of this part of journey. They have encouraged and supported me tremendously, and I will forever be grateful for that.

Never give up: 行到水穷处，坐看云起时

TABLE OF CONTENTS

3. QUANTITATIVE PROTEOMIC PROFILING STUDIES OF PANCREATIC CANCER STEM CELLS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Bottom-up Proteomics

The critical role of mass spectrometry (MS) in the field of proteomics has been firmly established. MS is a powerful tool to analyze gas phase ionized analytes based on their mass-to-charge (m/z) ratio. Proteomics involves studies which deal with large-scale profiling of protein expression levels, post-translational modification (PTM) levels and protein-protein interactions in a variety of complex biological systems. The marriage between MS and proteomics has made proteomics possible and widely broadened the application of MS-based techniques, leading to a continuous development of instrumentation to address the major challenges in proteomics. This includes such issues as the high degree of complexity of samples, and the masking of low abundance proteins. It is estimated that approximately 30,000 genes code for up to 30 times as many protein products with a concentration range varying by 10-12 orders of magnitude. The low abundance proteins which often exert important functions have lower signals due to ion suppression by high abundance proteins. Moreover, the interface between MS and proteomics has enabled the use of MS to solve important biological questions[1].

There are two major branches of MS-based proteomics strategies: top-down and bottom-up[2]. Top-down proteomics is defined as the analysis of intact proteins by MS while bottom-up is defined as the analysis of enzymatic digested proteins, often times tryptic peptides. These two approaches are complementary to each other and face the same degree of challenges. The strength of top-down lies in the direct detection of intact proteins so that the native primary structural information (such as PTMs, isoforms) is preserved. This method suffers from much less efficient ionization of large molecule as compared to peptides as well as the difficulty of sample handling and separation due to solubility issues. In the case of bottom-up proteomics, the well defined cleavage sites targeted by trypsin facilitate the bottom-up method dealing with peptides which are easier to solubilize and which can be separated with high-resolution HPLC and have improved ionization efficiency enabling the use of a wide range of mass spectrometers. The drawback of bottom-up proteomics is that the aforementioned native information is lost.

Bottom-up proteomics has become the method of choice in my thesis projects mainly because: 1) It is the best approach suited to the electrospray ionization (ESI)-ion trap mass spectrometer (Thermo LTQ, which has a superior sensitivity over other mass spectrometers) as compared to top-down; 2) PTM profiling is still possible by using an affinity chromatography method prior to MS; 3) Database and search engines are well developed to analyze bottom-up MS data.

To be noted, all the identification and quantification mentioned in this dissertation are performed at the peptide level.

1.2 Quantitative Strategies

In addition to the identification of proteins present within a system at a given time or under a particular perturbation condition, the quantification of protein expression levels or PTM levels is increasingly required because it can be viewed as a function of cellular state to infer molecular mechanisms. In particular, differential proteomic profiling which compares paired or multiple samples with biological relevance has become a principle strategy to identify biomarkers or to study the altered underlying signaling events.

1.2.1 Overview

Generally, the quantitative strategies for bottom-up proteomic research can be categorized into three groups: stable isotope labeling, label-free and multiple reaction monitoring (MRM).

1.2.1.1 Stable Isotope Labeling

The stable isotope labeling method is facilitated by the similar chemical and physical properties of the natural compounds and their isotope labeled counterparts except for the m/z. The quantification is performed by incorporating the isotope labeled molecules into MS analysis as internal standards or relative references[3]. A number of approaches have been developed under this category including Isotope-Coded Affinity Tag (ICAT)[4], Isobaric Tags for Relative and Absolute Quantification (iTRAQ)[5], Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)[6]. These methods have enabled simultaneous identification and quantification in complex samples. However, these approaches have potential drawbacks which limit the desired high through-put fashion of MS analysis such as: 1) increased analysis time from more complicated sample preparation procedures and data processing steps; 2) higher cost of required reagents; 3) incomplete labeling; 4) limited number of samples to be compared; 5) limited quantification dynamic range; 6) particular instrument requirements. The last one is especially critical in my thesis projects. It is the major factor limiting the use of these isotope labeled methods. The LTQ mass spectrometer, a major instrument used in the projects, is not compatible with most of the isotope labeling methods due to its “low mass cut-off” feature, meaning that the reporter ions used for quantifications at the MS2 level have poor signals under the commonly employed collision associated dissociation (CID) mode. The only alternative is to operate the LTQ under pulsed-Q-dissociation (PQD) mode, however, it is a tedious procedure and it has also already raised a question of quantification accuracy among the science community.

1.2.1.2 Label-free

The label-free method has become the preferred method of choice for quantification of global protein expressions due to the aforementioned concerns. By definition, label-free represents a strategy which avoids the isotope labeling step. There are two categories of label-free based measurements: peak area (or ion intensity) and spectral counting. These two methods are mostly used for relative quantification purpose. Peak area describes a method that measures the quantity of analytes based on the integrated peak area from the extracted ion chromatogram (EIC). The principle is that the detected ion signal is positively proportional to the analyte concentration by ESI within a certain range when coupling with liquid chromatography (LC). A typical data processing procedure of this kind of measurement includes peak detection (a particular peptide peak is distinguished from the noise background and neighboring peaks) and peak matching

(LC-MS) retention time is adjusted and the isotopic peaks are resolved for each peptide). It is obvious that this type of label-free method has some practical constraints. First, the LC-MS must be highly reproducible. Any drifts in retention time and m/z will significantly complicate the peak alignment process. Although a few computational algorithms have been developed to automatically perform the adjustments, the requirement of small experimental variations still holds. Second, the MS instruments used must be high- resolution; otherwise it creates a large obstacle for the peak detection process because of the ambiguity of distinguishing overlapping peaks by using low-resolution instruments. In addition, profile data containing the information regarding peak shapes rather than centroid data (peak lists) are acquired which further raises the bar for compatible computers with large storage capacity and high central processing units.

The other label-free method spectral counting represents a much simpler and straight-forward measurement strategy. Spectral counts are defined as the number of MS2 spectra assigned to one protein. Thus, it measures how many times the MS2 events are performed for one peptide selected in the MS1 stage and then sums up the number of MS2 events for each peptide belonging to one protein. This is based on the observation that there is a typically positive correlation between protein abundance and the number of its proteolytic peptides and vice versa[7]. Another more detailed study has shown that relative protein abundance is mostly correlated with spectral counts when compared to other factors such as sequence coverage and peptide number[8]. Interestingly, it was found that spectral counting and peak area exhibited strong correlation when EIC with high signal to noise (S/N) were used in the comparison. In addition, spectral counting is more reproducible and has a larger dynamic range[9]. It has been reported that spectral counting expands the dynamic range, allowing for the detection of abundance differences up to 60:1, whereas the ratio is 20:1 in SILAC experiments[10]. Thus, spectral counting has gained increasing popularity.

Another advantage of spectral counting over peak area lies in the zero software requirements, whereas peak area needs very sophisticated computational algorithms to handle the LC-MS data for feature detection and peak alignment. Spectral count information is embed in the MS/MS spectra searched results. Only several lines of codes are needed to parse out the spectral counts. Moreover, spectral counts are obtained from centroid data which means this method is suited to low resolution (unit mass) mass spectrometers such as the LTQ.

Therefore, the spectral counting based label-free method has been employed as the major quantitative tool in my thesis projects for a rapid screening to obtain a global view of the altered protein expression patterns in differential proteomic studies.

1.2.1.3 Multiple Reaction Monitoring

Selected Ion Monitoring (SIM) is a scan mode in ion-trap or triple quadrupole mass spectrometers with enhanced sensitivity and specificity since only a particular ion is selected to be analyzed at a given time. Multiple Reaction Monitoring (MRM) is another scan mode with even greater specificity and it is achieved by selecting a specific transition meaning that only a particular pair of parent ion and product ion is monitored at a given time[11]. Basically, any triple quadrupole type mass spectrometers is capable of performing such analysis including hybrid instruments such as the Q-TOF and Q-TRAP. MRM has been a principle tool for quantifying small molecules for decades[11]. The application of MRM in measuring the protein quantity based on the selected proteolytic peptides has just emerged in recent years and has undergone an increasing expansion. MRM represents the most accurate MS-based quantification strategy to date. This is because: 1) sensitivity is greatly increased since a dwell time is allocated for monitoring each MRM only; 2) specificity is also largely enhanced due to the monitoring of precursor ion and product ion simultaneously; 3) quantification accuracy is also optimized by either spiking a known amount of isotope labeled analog as the internal standard for absolute quantification or generating an external calibration curve by non-isotope labeled synthetically identical targeted peptide.

Because of the unique advantages of MRM, it is a preferred strategy in target verification. Conventionally, Enzyme-linked Immunosorbent Assay (ELISA) and Western Blot are the major validation methods. However, these techniques are not truly quantitative and the time to develop a new ELISA assay is a lengthy process. Nowadays, there has been a trend to develop a scheduled MRM (performing hundreds of MRM assays within a single LC run) as a complementary/alternative method for verification purpose. This is particularly advantageous since it could benefit from a combination of key advantages: high throughput, multiplex and accurate quantification.

Therefore, in addition to the traditional Western Blot method, MRM has been utilized as an alternative to confirm the fold changes detected by the spectral counting based label-free method in the global discovery phase.

1.2.2 Computational Challenges Associated with Spectral Counting

The peak area method has a requirement of special software to deal with the LC

MS data whereas spectral counting does not need any special software because of the ease of implementation[3]. However, there are still several challenges which are elaborated as follows.

1.2.2.1 Experimental Variance

Experimental variations are inevitable even when the same sample is analyzed in replicates by sequential LC-MS runs under exactly the same conditions. This is largely due to the random sampling nature of the data-dependent acquisition mechanism employed by the LTQ. Also, variations can be introduced in any sample handing step and by uncontrolled factors such as inherent instrumental drift. Thus, it is necessary to perform data normalization in the first place to minimize such variations. The normalization methods can be categorized as: 1) global normalization; 2) use of a housekeeping protein as a control; 3) spiking an internal standard as a control; 4) normalized spectral abundance factor (NSAF).

Global normalization is the simplest and most widely used approach which is also shared by the analysis of microarray data. The spectral counting data for each LC-MS run can be normalized against the total spectral counts[12], the mean, the median or to match the percentiles of each run to account for the variations. The abundance of housekeeping proteins such as Actin can also be used as a correction factor. The use of an internal standard such as spiking BSA into each sample is also facilitated to correct for the run-to-run variation[13]. Another normalization strategy called NSAF[14] is a more sophisticated method tailored to handle spectral counts, addressing the concern that proteins with longer length normally generate more tryptic peptides and have potentially more spectral counts. Although a variety of normalization methods are proposed, there is no universally recognized standard and the choice is rather empirical.

1.2.2.2 Assignment Ambiguity

Errors in the assignment of peptides to their corresponding protein directly propagate into protein abundance index using a spectral counting-based method. This occurs during the protein identification process where it is difficult to resolve proteins belonging to peptides which map to multiple protein sequences. Although methods have been developed to calculate the likelihood of MS2 peptide assignments[15], currently there is a lack of robust models to resolve such ambiguities of assembling peptides back to the protein[16]. Thus, this problem has a direct impact on the accuracy of the spectral counting method due to its dependence on the quality and quantity of MS2 spectra identifications.

1.2.2.3 Data Discontinuity

Another major challenge for spectral counting is that low abundant ions which are not selected for MS2 fragmentation will receive a spectral count equal to zero, termed “data discontinuity” or the “missing data” problem. Basically this is because of the random sampling nature of the LTQ where the selection of precursor masses for MS/MS analysis is skewed toward peptides of high abundance and the identification of low abundant peptides is a more random event, thus it is less reproducible[16]. This is illustrated in Figure-1. For example, the data-dependent acquisition is programmed to select the five most abundant ions at a given time. Setting up a dynamic exclusion window can compensate for this problem, but there is still a chance that the sixth ion is never selected for MS2 fragmentation during the entire LC-MS run. The chance is further increased as the abundance of this ion is decreased. In contrast, the peak area method does not have this problem because the sixth ion is still detected as long as it is above the S/N threshold. This phenomenon is attributed to the different fundamental quantification mechanism that spectral counting measures at the MS2 level where random sampling of low to medium abundance proteins occurs, whereas peak area measures at the MS1 level.

This random sampling problem of low to medium abundance proteins can be alleviated by repeating multiple replicate runs. Another way to correct for the missing data problem is by employing a computational strategy where data transformation is involved. Specifically, it is performed by using a correction factor to replace the missing value (spectral count = 0). Old *et al*[17] reported such a transformation strategy using a $\log_2$ scale quantity for each protein:

$$N= \log_2[(n+f)/(t-n+f)] \qquad \text{(Eq.1)}$$

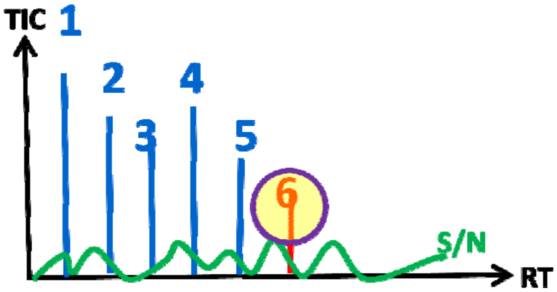


Figure 1.1: Illustration of the data discontinuity problem in spectral counting method caused by the data-dependent mechanism. The sixth most abundant ion is not selected for MS2 event so that it is assigned with a spectral count equal to zero, whereas this ion is still assigned with a certain value in peak area method as long as it is above the S/N threshold level.

where n is the raw or normalized spectral count value; t is the total number of spectra over all proteins in each dataset; and f is a correction factor. Several procedures for setting the constant term f have been proposed. A new approach having similar principles is devised and elaborated in Chapter3.

1.2.2.4 Differential Expression

Comparing spectral counting data from different stages or different sample groups is a critical step in statistical analysis. It is the most important data processing step as it answers the essential question in differential proteomic studies: which molecules are truly altered on their protein expression levels? To increase the confidence for determining if a molecule is significantly differentially expressed, different statistical tests have been employed and evaluated in various scenarios. The Student's t-test has been found to be the best performer when three or more replicates are available, while the Fisher's exact test, G-test and AC test are more suitable when the number of replicates is less than three[18].

However, determining which statistical test to choose is never a simple question due to the complexity of the spectral counting-based MS data. The task in reality is extraordinarily challenging. First, it remains uncertain which distribution fits best to the MS data. The assumption of t-test is actually violated as most MS data points are not normally distributed. This can be simply tested by plotting the sample quintiles against the theoretical quintiles (Normal Q-Q plot) along with quite a few statistical tests. Therefore, other distribution models have been explored, specifically Poisson distribution[19] and Beta-Binomial distribution[20] have been proposed to model the spectral counting data by other groups. Second, the prerequisites of many data analysis tools are that individual molecules are statistically independent. However, proteins are biological building blocks and are inherently correlated. Currently, few studies[21] have been reported to tackle this problem. Computational models which can adequately address this critical dependency issue still remain to be established. Third, the multivariate models have not been explored explicitly, especially in the case of seeking for a robust handling of hierarchical data structure. Figure-2 illustrates such a structure where there are two sample groups with two levels of replicates/variances. The shortcoming of the commonly employed tests such as the t-test is that it does not take into account within- and between- sample variations together.

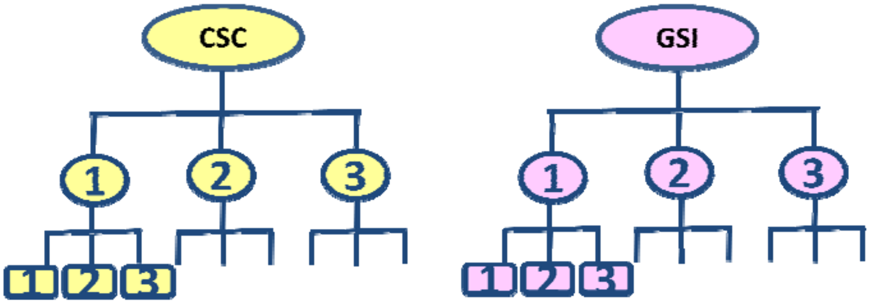


Figure 1.2: A typical hierarchical data structure. CSC represents a control group and GSI represents a treatment group. Each group has three biological replicates depicted in circles and each biological replicate has three technical replicates depicted in squares.

Therefore, the generalized linear mixed effect model (GLMM)[22] which alleviates the above three issues to some degree is a better strategy to test the significance of differential protein abundances. Specifically, the between-group difference is modeled as fixed effect and the within-group difference is modeled as random effect using the restricted maximum likelihood (REML)[23] procedure to estimate the parameters. The use of this model is described in Chapter4.

1.3 Dissertation Outline

This dissertation consists of four chapters elaborating four different but interconnected bottom-up proteomic projects. Chapter2 describes a comparative proteomic profiling study of two closely related ovarian cell lines. A two-dimensional separation by coupling cIEF and reversed phase LC (RPLC) is utilized together with the LTQ mass spectrometer. Pathway analysis is performed to infer altered signaling events. This first project has paved the way for the following projects which are concentrated on applying the principal of the established methodology to study cancer stem cells (CSCs). CSCs are termed by their unique properties and are suggested as novel and potentially revolutionized therapeutic targets. Specifically, Chapter3 is an extension of Chapter2 by further exploring and optimizing the cIEF technique to address the technical difficulty of extremely small sample quantity in the study of pancreatic CSCs. The data discontinuity problem is also addressed by utilizing a transformation strategy. Chapter4 describes the investigation of proteomic research in Glioblastoma Multiforme (GBM) CSCs upon Gamma Secretase Inhibitor (GSI) treatment in order to obtain a better understanding of drug impact. Instead of profiling the protein expressions at a global view, altered glycosylation levels between the GBM CSC group and the drug treatment group are interrogated by coupling an affinity chromatography method prior to MS analysis. The differential expression by pairwise t-test and GLMM are both discussed. Chapter5 discusses a more comprehensive bottom-up proteomic study compared to the previous three projects. A complete pipeline consisting of global discovery by label-free quantification, candidate prioritization and target verification by MRM is described. The biological setting is similar to Chapter4. However, the question is addressed from a dose-dependent point of view. Also, biological implications are discussed by mapping important proteins into relevant signaling pathways to infer the altered signaling events upon drug treatment.

1.4 References

1. Aebersold R, Mann M, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.
2. Chait B.T, *Mass spectrometry: bottom-up or top-down?* Science, 2006. **314**(5796): p. 109-12.
3. Zhu W, S.J., Huang CM., *Mass spectrometry-based label-free quantitative proteomics.* J Biomed Biotechnol, 2010.
4. Smolka MB, Z.H., Purkayastha S, Aebersold R., *Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis.* Anal Biochem, 2001. **297**(1): p. 25-31.
5. Ross PL, H.Y., Marchese JN, 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 DJ., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.
6. Everley PA, K.J., Zetter BR, Gygi SP., *Everley PA, Krijgsveld J, Zetter BR, Gygi SP.* Mol Cell Proteomics, 2004. **3**(7): p. 729-35.
7. Liu H, S.R., Yates JR 3rd., *A model for random sampling and estimation of relative protein abundance in shotgun proteomics.* Anal Chem, 2004. **76**(14): p. 4193-201.
8. Washburn MP, W.D., Yates JR 3rd., *Large-scale analysis of the yeast proteome by multidimensional protein identification technology.* Nat Biotechnol, 2001. **19**(3): p. 242-7.
9. Zybailov B, C.M., Florens L, Washburn MP., *Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling.* Anal Chem, 2005. **77**(19): p. 6218-24.
10. Asara JM, C.H., Freimark LM, Cantley LC., *A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen.* Proteomics, 2008. **8**(5): p. 994-9.
11. Kitteringham NR, J.R., Lane CS, Elliott VL, Park BK., *Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics.* J Chromatogr B Analyt Technol Biomed Life Sci, 2009. **877**(13): p. 1229-39.
12. Dong MQ, V.J., Au N, Xu T, Park SK, Cociorva D, Johnson JR, Dillin A, Yates JR 3rd., *Quantitative mass spectrometry identifies insulin signaling targets in C. elegans.* Science, 2007. **317**(5838): p. 660-3.

13. Stevenson SE, C.Y., Ozias-Akins P, Thelen JJ., *Validation of gel-free, label-free quantitative proteomics approaches: applications for seed allergen profiling.* J Proteomics, 2009. **72**(3): p. 55-66.

14. Paoletti AC, P.T., Tomomori-Sato C, Sato S, Zhu D, Conaway RC, Conaway JW, Florens L, Washburn MP., *Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors.* Proc Natl Acad Sci U S A, 2006. **103**(50): p. 18928-33.

15. Nesvizhskii AI, K.A., Kolker E, Aebersold R., *A statistical model for identifying proteins by tandem mass spectrometry.* Anal Chem, 2003. **75**(17): p. 4646-58.

16. Mueller LN, B.M., Mani DR, Aebersold R., *An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data.* J Proteome Res, 2008. **7**(1): p. 51-61.

17. Old WM, M.-A.K., Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG., *Comparison of label-free methods for quantifying human proteins by shotgun proteomics.* Mol Cell Proteomics, 2005. **4**(10): p. 1487-502.

18. Zhang B, V.N., Langston MA, Uberbacher E, Hettich RL, Samatova NF., *Detecting differential and correlated protein expression in label-free shotgun proteomics.* J Proteome Res, 2006. **5**(11): p. 2909-18.

19. Choi H, F.D., Nesvizhskii AI., *Significance analysis of spectral count data in label-free shotgun proteomics.* Mol Cell Proteomics, 2008. **7**(12): p. 2373.

20. Pham TV, P.S., Warmoes M, Jimenez CR., *On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics.* Bioinformatics, 2010. **26**(3): p. 363-9.

21. Choi H, L.B., Lin ZY, Breitkreutz A, Mellacheruvu D, Fermin D, Qin ZS, Tyers M, Gingras AC, Nesvizhskii AI., *SAINT: probabilistic scoring of affinity purification-mass spectrometry data.* Nat Methods, 2010.

22. Zeger, S.L., Karim, M. R., *Generalized linear models with random effects; a Gibbs sampling approach.* J. Am. Stat. Assoc, 1991. **86**(413): p. 79-86.

23. Breslow NE, C.D., *Approximate Inference in Generalized Linear Mixed Models.* J. Am. Stat. Assoc, 1993. **88**(421): p. 9-25.

CHAPTER 2

COMPARATIVE PROTEOMIC STUDIES OF OVARIAN CANCER CELLS

2.1 Abstract

The proteomic profiles from two distinct ovarian endometrioid tumor-derived cell lines, (MDAH-2774 and TOV-112D) each with different morphological characteristics and genetic mutations, have been studied. Characterization of the differential global protein expression between these two cell lines has important implications for the understanding of the pathogenesis of ovarian endometrioid carcinoma. In this comparative proteomic study, extensive fractionation of peptides generated from whole-cell trypsin digestion was achieved by coupling cIEF in the first-dimensional separation with capillary LC (RP-HPLC) in the second dimensional separation. Online analysis was performed using tandem mass spectra acquired by a linear ion trap mass spectrometer from triplicate runs. A total of 1749 and 1955 proteins with protein probability above 0.95 were identified from MDAH-2774 and TOV-112D after filtering through Peptide Prophet/Protein Prophet software. Differentially expressed proteins were further investigated by ingenuity pathway analysis (IPA) to reveal the association with important biological functions. Canonical pathway analysis using IPA demonstrates that important signaling pathways are highly associated with one of these two cell lines versus the other, such as the PI3K/AKT pathway, which is found to be significantly predominant in MDAH-2774 but not in TOV-112D. Also, protein network analysis using IPA highlights

p53 as a central hub relating to other proteins from the connectivity map. These results illustrate the utility of high throughput proteomics methods using large-scale proteome profiling combined with bioinformatics tools to identify differential signaling pathways, thus contributing to the understanding of mechanisms of deregulation in neoplastic cells.

2.2 Introduction

Ovarian cancer is the fifth leading cause of cancer-related death in the Western world and causes more deaths of women in the United States than any other gynecological malignancy [1]. The five-year survival rate can be as high as 90% with early detection; however, early detection of ovarian cancer is rare and known markers have limited utility for general population screening. The most common form of ovarian cancer is epithelial ovarian cancer, which can be further divided into four major histological subtypes: serous, clear cell, mucinous and endometrioid [2]. Ovarian endometrioid adenocarcinoma (OEA) represents approximately 20% of common epithelial tumors.

In the present comparative study, we have employed two closely related OEA cell lines, MDAH-2774 and TOV-112D[3,4]. Both of these two cell lines were derived from female Caucasian patients with OEA. In particular, the TOV-112D cell line originates from an aggressive ovarian endometrioid tumor (stage 3, grade 3). The growth characteristics and tumorigenic potential of this cell line parallels the clinical behavior of aggressive OEAs. Categorization of the tumor grade/stage of MDAH-2774 is not available. Differential global gene expression analyses have been performed, and different genetic defects have been previously detected between these two cell lines, possibly leading to different levels of deregulation of important signaling pathways[5]. It has been shown that both the MDAH-2774 and TOV-112D cell lines have elevated constitutive Wnt signaling deregulation. A missense AXIN1 sequence alteration was identified in MDAH-2774 and mutant beta-catenin was identified in TOV-112D. A mutated K-ras gene, involved in the PI3K/AKT signaling pathway, was detected in MDAH-2774 but not in TOV-112D. Both the MDAH-2774 and the TOV-112D OEA cell lines have a mutant p53 gene[6].

Protein expression and gene expression data, while being mutually exclusive, are complimentary to each other. A lack of direct correlation between protein expression and gene expression has been reported [7]. Protein over/under expression is expected to relate to deregulated tumor cell behavior more directly than would gene expression. The proteomic profiles of these two cell lines have been generated in previous work using different methods. In one study, Rotofor IEF and nonporous (NPS) reversed phase separation was coupled with ESI-TOF-MS and MALDI-TOF-MS to analyze the proteome of MDAH-2774 via intact protein fractionation [8]. In a second study, 2D-PAGE coupled to MALDI-TOF-MS and SDS-PAGE coupled to LC-MS/MS were both used to obtain protein profiles from TOV-112D [9]. Alternatively, a shotgun [10] proteomics strategy of a whole cell digest can be used to compare the global proteome profile of MDAH-2774 and TOV-112D (both qualitatively and quantitatively) in order to analyze protein expression differences in neoplastic dedifferentiation. Within, we have utilized capillary isoelectric focusing (cIEF) to separate peptides based on pH[11-12], followed by capillary reversed phase separation with on-line nanoESI-ion trap mass spectrometer analysis. This method is capable of identifying large numbers of proteins over an extended pH range where 1749 and 1955 proteins from triplicate runs of MDAH-2774 and TOV-112D, respectively, have been identified in this work.

Quantitation is always an important issue in pathway analysis using either isotopic labeling or label free methods. Label-free quantitation has gained increasing popularity in recent years and has been successfully applied in large quantitative studies [13-14] due to the development of computational and statistical methods and advances in LC-MS/MS systems. Extraction of peptide ion intensities and spectral counting (defined as the number of MS/MS spectra identified per protein) are two widely adopted methods for performing comparative quantitative analysis of LC-MS proteomics experiments. It has been shown that spectral counting is highly reproducible and is sensitive to protein abundance changes[15]. Further, in controlled experiments it was found that the correlation of protein abundance with spectral count is superior to that of protein sequence coverage or peptide count[14]. Thus, we have utilized spectral counting to measure protein abundance. The ratio of the spectral count of the same protein represents the relative expression level between two samples. Spectral sampling can enable protein ratios larger than ~2-fold to be determined with high confidence.

The large number of identified proteins between these two cell lines provides a means for qualitative and quantitative bioinformatics pathway analysis. Differentially expressed proteins can be further investigated to reveal the associated biological pathways using bioinformatic tools such as Ingenuity Pathway Analysis (IPA). The IPA program uses a knowledge base derived from the literature containing information on interactions between genes, proteins and other biological molecules. After uploading differentially expressed protein lists to the IPA server, IPA uses these focused proteins to extract connectivity networks which relate candidate proteins to each other based on their interactions and generates global canonical pathways which are shown to be significantly associated with these candidates[16]. As illustrated in Figure 2.1, we have used a strategy of shotgun proteomics with subsequent bioinformatics analysis to study pathways in the TOV-112D and MDAH-2774 cell lines in order to understand the different interactions in these two OEA cell lines.

2.3 Materials and Methods

2.3.1 Sample preparation

MDAH-2774 and TOV-112D cells were gently washed 3 times with PBS (pH 7.4) by repetitive pipetting, followed each time by centrifugation at 1,500×g for 5min at 4 C. The cell pellets were resuspended with 1 ml lysis-denaturing buffer (7.5M urea, 2.5M thiourea, 12.5 v/v glycerol, 62.5M Tris-HCl, 2.5%(w/v) n-octylglucoside (n-OG) and 1% v/v protease inhibitor cocktail). All chemicals were purchased from Sigma (St. Louis, MO) unless otherwise noted. The lysates were vortexed and then centrifuged at 35,000×g for 1hr at 4 C. The supernatant was collected and dialyzed against 50 mM ammonia bicarbonate overnight using Slide-A-Lyzer dialysis cassettes with a 3,500 Da molecular cutoff from Pierce (Rockford, IL). The proteins were quantified with the micro-BCA assay kit from Pierce (Rockford, IL), and then lyophilized to 100 μl using a SpeedVac concentrator (Labconco, Kansas City, MO) operating at 45 C.

2.3.2 Trypsin Digestion

5mM dithiothreitol (DTT) was added and the mixture was incubated at 60 C for 30min. After cooling, 5mM iodoacetamide (IAA) was added and the mixture was placed in the dark at room temperature for 30 min in order to carboxamidomethylate the Cysteine residues. Then, 1:50 w/v L-1-tosylamido-2-phenylethyl chloromethylketone (TPCK) modified sequencing-grade porcine trypsin from Promega (Madison, WI) was added. Following vortexing, the mixture was incubated overnight at 37°C in a water bath with agitation, followed by addition of 2% formic acid (FA) to terminate the reaction.

2.3.3 First dimension separation: cIEF

Peptides were sequentially resolved based on their different isoelectric points (pI) and hydrophobicity. cIEF was performed on a Beckman CE instrument with sample collection as shown in Figure 2.2. A 70cm cIEF (100um i.d. 365um i.d.) capillary was coated with hydropropyl cellulose for eliminating electroosmotic flow and absorption of peptides onto the capillary wall. The capillary was initially filled with sample gel buffer containing 2% ampholyte 3-10 and 1μg tryptic peptides. Sodium hydroxide solution at pH 10.8 and 0.1M phosphate acid solution were employed as catholyte and anolyte, respectively. One end of the capillary was emerged in the anolyte, while the other end was kept in coaxial metal tubing with a sheath flow composed of catholyte eluting flush with the exit of the capillary. The flow rate was controlled by a syringe pump at 2 μl/min, and was adjusted to ensure that a proper droplet formed at the exit to carry the peptides fractionated into individual wells in the sample plate. Isoelectric focusing was performed at 21kV (300V/cm) over the entire capillary. The current decreased continuously as the peptides were focused and the process was considered complete after the current no longer changed. The focused bands of peptides were sequentially mobilized slowly under pressure towards the cathode and delivered as droplets with catholyte sheath flow into individual wells on a sample plate, where the fractions were collected with a modified

Beckman HPLC sample collector. Each cIEF separation runs approximately 90 min. One-third of the run time is spent in focusing the peptides in the capillary while the remaining time is used to deposit the off-line fractions.

2.3.4 Second dimensional separation: nanoRPLC+nanoESI-MS/MS

When cIEF separation was completed, each pI fraction of tryptic-digested sample was injected via Paradigm autosampler (Michrom Biosciences, Auburn, CA) and loaded onto a desalting Nano trap (150µm×50mm) (Michrom) connected to a nano RP column(C18AQ, 5µm 200A 0.1×150mm) (Michrom) by a Paradigm AS1 micropump (Michrom). The mobile phase A and B were composed of 0.3% FA in water and 0.3% FA in acetonitrile (ACN), respectively. Peptides were first desalted and enriched starting at 100%A with a flow rate of 50 µl/min for 5 min. Sample was subsequently separated by a Nano RP column with a flow rate of 0.3 µl/min after splitting. The linear gradient for separation was as follows: from 3% ACN to 12% ACN in 5 min, from 12% ACN to 40% ACN in 30 min, from 40% ACN to 80% ACN in 15 min and finally decreased from 80% ACN to 3% ACN in 10 min. The resolved peptides were then introduced into a ThermoFinnigan LTQ mass spectrometer (Thermo Electron Corp., San Jose, CA) equipped with a nanospray ion source (Thermo). The LTQ was operated in data-dependent mode in which one cycle of experiments consisting of one full MS scan was followed by 5 pairs of zoom scans and MS/MS scans with dynamic exclusion set to 30 sec. The capillary temperature was set at 175 C, spray voltage was 2.8kV, capillary voltage was 30V and the normalized collision energy was 35% for the fragmentation.

2.3.5 Database Search and Protein Identification

MS/MS spectra were then searched against the human UniProt FASTA database (updated in Dec.2007) by TurboSEQUEST provided by Bioworks ver3.1 SR1 (ThermoFinnigan). The search was performed using the following parameters: (1) Enzyme: trypsin; (2) one missed cleavage allowed; (3) peptide ion mass tolerance: 1.5Da; (4) fragment ion mass tolerance: 1.0 Da; (5) mass tolerance for precursor ions 1.4Da; (6) peptide charges +1, +2, +3; (7) possible modifications: 15.99 Da shift for oxidized Met residues; 79.97 Da for phosphorylated Ser, Thr, Tyr residues respectively; 58.1 Da shift for carboxymethylated Cys residues. The identified peptides were subsequently processed through Peptide Prophet and Protein Prophet incorporated in the Trans Proteomic Pipeline (TPP)[17]. In TPP, the search results were first validated by Peptide Prophet, which converts various SEQUEST parameters to a discriminant score and uses Bayesian statistics to compute the probability that each identified peptide is correct. Protein Prophet reads in peptides and assigned probabilities to compute the probabilities of proteins that are present in the original sample (http://proteinprophet.sourceforge.net/prot-software.html). In this study, we use a protein probability score of ≥0.95 as the threshold for protein identification, to ensure that the minimized overall error rate is below 0.05.

2.3.6 Label-free Quantitation and Normalization

After processing the Sequest data through TPP, the spectral counts were parsed out of TPP protXML files using a perl script(see Figure 2.1b). Three datasets of identified proteins with 0.95 protein probability and their associated spectral count have been generated for each sample. We divided the data into two groups. Qualitative data consists of proteins that are only identified in one of these two cell lines, whereas quantitative data consists of proteins that are identified in both of these two cell lines with their expression values. In the first group one cannot compare the same protein expression level between two cell lines. In the second group the relative protein abundance fold change of the same protein can be calculated by the ratio of their spectral count in two samples as explained later in this work. The data was processed in two different ways. For the qualitative analysis, only the qualitative data was used (i.e. only different protein names from two samples were included). We also performed a quantitative analysis in which we combined the qualitative data and quantitative data by replacing any missing value with zero. For example, protein "A" is only detected in MDAH-2774 with its assigned spectral count. In order to make the comparison of the differential expression of this protein plausible, we assume protein "A" is also present in TOV-112D but at a very low level which is not detectable and assign a spectral count of zero to protein "A" in TOV-112D.

Subsequent normalization was used to reduce technical bias when acquiring spectral count data from different runs across the two different cell lines. The bias may come from instrument error or the inherent random sampling nature of the LTQ. In order to normalize the data, we first calculated the ratio of the total spectral count of 3 runs between MDAH-2774 and TOV-112D and then multiplied the spectral count of each protein in the numerator by this ratio. Statistical significance levels of the pair-wise comparison were then adjusted for multiple testing using the false discovery rate (FDR), q-value method. Differentially expressed genes used to learn network structure were declared at a FDR q-value threshold of 0.3. The FDR q-values were calculated using the R package[18].

2.3.7 Ingenuity Pathway Analysis

To infer global network functions between all differentially expressed proteins from MDAH-2774 and TOV-112D, we conducted two types of analysis using Ingenuity Pathway Analysis software (IPA). For the qualitative analysis, we uploaded into the IPA database two sets of proteins with corresponding primary accession number which were only identified from one of these two cell lines. Out of 652 and 838 proteins uploaded from MDAH-2774 and TOV-112D, the IPA software identified 515 and 665 "focus genes" that were eligible for generating connectivity networks and 443 and 582 "focus genes" that were eligible for generating biological functions/disease and associated pathways.

In order to gain further insight into the dynamic changes of the cell states between these two cell lines at the molecular level, we performed a quantitative analysis by incorporating quantitative data in addition to qualitative data. In this analysis, we uploaded a list of 828 differentially expressed proteins with fold change larger than 2 based on normalized spectral count data. The relevant proteins with their fold change, qv-value and corresponding primary accession number were uploaded as an Excel spreadsheet file. 609 proteins were eligible for generating networks and 532 proteins were used to retrieve functions/pathways after applying a threshold of qv-value of <0.3.

The significance values for analyses of network and pathway generation were calculated using the right-tailed Fisher's Exact Test by comparing the number of proteins that participate in a given function or pathway relative to the total number of occurrences of these proteins in all functional/pathway annotations stored in the Ingenuity Pathway Knowledge Base (IPKB).

2.3.8 Western Blot Analysis

MDAH-2774 and TOV-112D cell lines were lysed in lysis buffer as described above. 100 µg of total protein from each of the cell lysates was separated by 10% SDS-PAGE in parallel. The resolved proteins were transferred to PVDF membranes (Immobilon-P, Millipore) by conventional procedures using a TE70 semi-dry transfer unit (Amersham Biosciences, Princeton, NJ). Beta-actin protein expression levels were used as an internal control to ensure equal loading between lanes. After transfer, membranes were incubated with a blocking buffer consisting of PBS and 0.1% Tween 20 containing 5% nonfat dry milk overnight. The membranes were incubated for 1 h at room temperature with primary antibodies against UCHL1 (rabbit polyclonal antibody, Biogenesis, NH), SFN (mouse monoclonal antibody; Abcam, Cambridge, MA), MARKS (mouse monoclonal antibody, (Abcam) and beta-actin (mouse monoclonal antibody, (Sigma, St. Louis, MO)) for 1hr at 1:5000 dilution in Tris-buffered saline. Membranes were simultaneously incubated with the mouse anti-beta actin antibody and either the rabbit anti-UCHL1 antibody, the mouse anti-SFN antibody or the mouse anti-MARKS antibody. After three washes with washing buffer (PBS containing 0.1% Tween 20), the membranes were incubated with the appropriate secondary antibody (highly cross absorbed HRP-conjugated goat anti mouse and/or highly cross absorbed HRP-conjugated goat anti rabbit; Abcam) for 1hr at 1:2000 dilution. Immunodetection was accomplished by enhanced chemiluminescence (Amersham Biosciences) followed by autoradiography on Hyperfilm MP (Amersham Biosciences).

2.4 Results and Discussion

2.4.1 cIEF Performance

20ug of whole cell lysate was extracted from MDAH-2774 and TOV-112D followed by trypsin digestion. Each aliquot of tryptic peptides (~5ug) was then loaded to cIEF separation. About 40 fractions were collected per run and each fraction was further subjected to the second dimensional separation coupled with nanoESI-ion trap. We have repeated this process three times for each cell line sample.

The theoretical pI value for each identified peptide within each fraction was calculated after database searching. The pI distribution plot from the second run of MDAH-2774 is shown in Figure 2.3. As expected, the pI trend follows a non-perfect linear velocity. Peptides with pI in the region 3.5 to 8 tend to show improved separation performance compared to basic peptides with pI from 8 to 10. Peptides with pI above 10 or below 3 are not expected to be resolved as they fall outside the pH range of the ampholytes (pH 3-10) used in these experiments. Overall, cIEF exhibits high separation resolution with little overlap of the same peptides identified between adjacent fractions.

An important issue here is the use of offline collection of cIEF fractions coupled to nano-RPLC. With the use of on-line cIEF coupled to nano-RPLC one can directly load each fraction to a nano RP-column by sacrificing the separation resolution due to the transfer of cIEF fractions to the 2nd dimension from the increased back pressure and dead volume. In the offline collection method, the sheath-flow eluting from the coaxial tubing was adjusted flush with the exit of the capillary in order to eliminate back pressure and dead volume as shown in Figure 2.2. Compared to the online integration of cIEF/nano-RPLC, the offline collection mode does not degrade the cIEF separation, significantly reducing the mixing of separated peptides during the transfer process. This is also central for precise quantification by spectral counting in this work.

2.4.2 Proteomic Profiling

2.4.2.1 Number of proteins identified

MS/MS spectra were searched against the UniProt database using SEQUEST software and search results were then validated using the Peptide Prophet program. Peptide Prophet provides an empirical statistical model that estimates the accuracy of peptide identifications made by SEQUEST. For each tandem mass spectrum, Peptide Prophet determines the probability that the spectrum is correctly assigned to a peptide based upon its SEQUEST scores. A second program, Protein Prophet was subsequently used to group peptides by their corresponding proteins to compute probabilities that those proteins were present in the original sample. A stringent cutoff of 0.95 was used to filter all the SEQUEST results based on Protein Prophet's estimate of error rate. For each cell line, we have repeated the same experimental procedure and combined the results from all three runs instead of selecting only the overlapping proteins. This is done since some proteins can only be identified in a single run of a sample due to the random sampling nature of tandem MS. The Venn diagram in Figure 2.4a summarizes the intersection of proteins identified from all three runs of MDAH-2774. In the first run of MDAH-2774, 656 distinct proteins were identified from 25 fractions. 1181 and 1095 distinct proteins were identified in the second run and the third run of MDAH-2774, respectively, when we increased the fractionation number to approximately 40. The total number of proteins from the combined list is 1749 for MDAH-2774 and 1955 for TOV-112D with an overlap of 1092 as shown in Figure 2.4b.

2.4.2.2 Cellular Localization

Each identified protein was assigned a cellular localization based on information from the Swiss-Prot, Entrez Gene, and Genome Ontology (GO) databases. Figure 2.5 shows the cellular distribution of 1749 identified proteins from MDAH-2774 and 1092 identified proteins from TOV-112D. The majority are cytoplasmic and nuclear proteins for both of these two cell lines. Membrane proteins only occupy 6% of each total proteome, which is not surprising since the protein extraction method used in this study is not optimized for hydrophobic proteins.

2.4.3 Label-free Quantitation

Detecting protein quantity and the changes in this quantity between various stages or different samples is central to understanding the molecular process of the cell. We used the spectral count as the measurement of relative protein abundance because it has been shown to accurately reflect relative protein abundance with a linear correlation of over two orders of magnitude of dynamic range [15]. Spectral count was assigned to each identified protein followed by normalization and log transformation. The signal distribution in Figure 2.6a shows that the ratio of protein expression level between MDAH-2774 and TOV-112D follows a symmetric distribution. These two cell lines have approximately an equal number of proteins that are up-regulated or down-regulated when compared to each other. Only proteins with fold change larger than 2 between MDAH-2774 and TOV-112D, which is equal to 1 in $\log_2$ scale, are shown in Figure 2.6a and were used for further comparative analysis. About two thirds of the identified proteins fall into the column with fold change range between 2 to 4. The rest of proteins fall into a fold change from 4 to 32, with a few exceptions over 2 orders of magnitude. The distribution of q-values for all proteins regardless of the expression fold change is also plotted in Figure 2.6b.

Pearson Correlation Analyses have also been applied to assess the reproducibility and quality of the quantitative data. The first run of MDAH termed MDAH1 and the first run of TOV termed TOV1 are slightly experimentally different than the rest of runs in terms of the number of fractionations in the cIEF separation step. It would therefore make more sense to evaluate the reproducibility between (MDAH2 and MDAH3), (TOV2 and TOV3), which results in a moderate to high Pearson Correlation Coefficient of 0.75 and 0.85. Pearson Correlation Analyses have also been applied to any of the two runs from two distinct cell line samples. Results are summarized in Table 2.1.

Table 2.2 lists the top10 most abundant proteins in MDAH-2774 and in TOV-112D and the top10 most differentially expressed proteins based on the ratio of their spectral count from MDAH-2774 over TOV-112D. From the quantitation list, we observed that the most abundant proteins from both of these two cell lines are proteins related to structural elements like vimentin, actin and tubulin, as well as chaperone proteins and members of the heat shock protein family. Proteins that are most differentially expressed between these two cell lines cover a wide range of molecular functions and associations with different diseases. For example, collagen3 alpha1 is a structural constituent of intracellular matrix. Tubulin beta4 is the major constituent of the microtubules. They have both been shown to be associated with epithelial ovarian cancer. Stratifin, a protein kinase C inhibitor, is involved in regulation of progression through the cell cycle and has been shown to be associated with breast cancer and prostate cancer. Eukaryotic translation elongation factor 1 alpha 2 has been shown to be associated with breast cancer.

Myristoylated alanine-rich protein kinase C substrate has been shown to be associated with endometriosis.

Other differentially expressed proteins that are not shown in this table also have important implications on the mechanisms of ovarian endometrioid adenocarcinoma (OEA). For example, beta-catenin (CTNNB1), a critical component of the Wnt signaling pathway, was found to be over-expressed 4.2-fold in TOV-112D as compared to MDAH-2774 based on our spectral count data. This compares favorably with previously reported data by Wu *et al.* in which CTNNB1 was expressed 4.4-fold in TOV-112D over MDAH-2774 from the CTNNB1/TCF transcription reporter assay [5]. Although CTNNB1/TCF transcriptional activity in MDAH-2774 is modest compared with TOV-112D, it is known to be present at elevated levels in both these two cell lines compared to other ovarian cell lines leading to constitutive activation of the Wnt signaling pathway. Notably, the CTNNB1 missense mutation was detected in TOV-112D by PCR sequencing [5]. It has a mutation in its NH_2-terminal regulatory domain, thereby rendering the mutant protein resistant to degradation thus resulting in a higher CTNNB1 level in TOV-112D than in MDAH-2774.

The most significant drawback of spectral counting is that it is more likely to be influenced by the acquisition program of the mass spectrometer compared to other label free comparative quantitation methods such as peptide ion intensity-based quantification. High abundance peptides can mask low abundant peptides if the data dependent MS/MS acquisition exclusion list is too small. If the exclusion list is too large, the spectral count can become rapidly saturated, resulting in reduced sensitivity. We have optimized the conditions in this case by extensive fractionation and setting the exclusion list time to 30 sec.

2.4.4 Comparison with Previously Reported Proteins

The proteome of both MDAH-2774 and TOV-112D cell lines have been previously analyzed by different methods. In the first study, a 2D all liquid phase (Rotofor IEF nonporous silica (NPS) RP-HPLC) separation method was used combined with ESI-TOF-MS and MALDI-MS/MS to compare the proteome profile of cultured ovarian cancer cell lines[8]. In this study, 161 unique proteins from MDAH-2774 were identified from five fractions with pH range from 5.8 to 8.3 by using PMF and peptide sequencing analysis after applying a 0.95 probability from the Mascot Search Engine. Around 70% of the proteins identified in the first study were also observed in the current study, including some important cancer-associated proteins such as the Oncoprotein 18/stathmin, ezrin and p53 protein. Oncoprotein 18/stathmin, a conserved cytosolic phosphoprotein that regulates microtubule dynamics, was identified in two of the three runs of the MDAH-2774 cell line. It was previously reported that over-expression of OP18 is associated with a variety of human cancers, including breast cancer and lung cancer[19,20]. Ezrin is a member of the ezrin/radixin /moesin family of membrane-axin cross-linking proteins that also transduces signals from growth factors. Previous studies have shown frequent ezrin over-expression in ovarian carcinomas, particularly in metastatic lesions[21]. Mutant P53 is also known to be over-expressed in MDAH-2774[22].

In another publication [9], the proteome profiling of TOV-112D has been examined by two complementary proteomic approaches, two-dimensional gel electrophoresis (2D PAGE) protein separation coupled to MALDI-TOF/MS and SDS-PAGE coupled to LC-MS/MS. 172 proteins were identified from 2D PAGE and a total of 589 proteins were identified from SDS-PAGE LC-MS/MS after applying a 0.9 probability cutoff by Protein Prophet, of which 436 proteins are also found in the current study. Relatively high expression of stress proteins like HSP90 and HSP71 were observed when compared to other proteins in both studies, as well as in numerous malignant tumors [23]. Two forms of aldehyde dehydrogenase1 which have previously been shown to be over-expressed in aggressive EOC versus non-aggressive EOC or normal ovarian surface epithelia cells at the RNA level were also observed in both studies[24]. Proteins that have been previously proposed as biomarkers or targets for diagnostic studies of invasive ovarian cancer because of their over-expression in invasive carcinomas as compared with benign tumors have been identified in previous studies[25] including FK506-binding protein 4 and several reported differentially expressed proteins such as proliferating cell nuclear antigen (PCNA); leukemia-associated phosphoprotein (stathmin); glutathione S-transferase π (GST π); triose-phosphate isomerase (TPI) and tumor metastatic process-associated protein (Nm23), which have been the subject of extensive investigation in ovarian cancer. In addition, Cytokeratin 18 and Cytokeratin 8 reported as biomarkers by Alaiya *et al.* [26]were also identified in our study, but not in the work of reference [9].

Overall, more than 70% of the proteins identified in previous work [8,9] were also found in our study when comparing our proteomic profiling results to previously reported data. Coupling of off-line cIEF with online nano-RPLC and nano-ESI-LTQ in our study has enabled a more comprehensive proteomic profiling of differentially expressed proteins between MDAH-2774 and TOV-112D.

2.4.5 Ingenuity Pathway Analysis

2.4.5.1 Qualitative Analysis

The 15 most variant canonical signaling pathways between these two cell lines were generated by IPA and are shown in Figure 2.7 with a threshold p-value<0.1 indicated. The length of the bar only indicates that the differentially expressed proteins are related to this pathway, but is by no means indicative of the pathway being either up-regulated or down-regulated. It is possible that the overall activity of a pathway is up-regulated or down-regulated, but it is not sufficient to draw a conclusion of the direction of change based on the data forming network alone. It is shown that MDAH-2774 and TOV-112D have different levels of association with different signaling pathways. For example, PI3K/AKT signaling was found to be more significant in MDAH-2774 than in TOV-112D from this figure. Previous studies have shown that frequent activation and over-expression of PI3K are associated with ovarian carcinoma[27]. Specifically, amplification of the catalytic subunit alpha of PI3K (PIK3CA) is detected in most ovarian cancer cell lines and primary tumors, as well as the somatic mutations in the gene encoding the p85α regulatory subunit of PI3K (PIK3R1) which leads to constitutive activation of PI3K. PIK3R1 was identified from MDAH-2774 but not in TOV-112D in this study, implying that PI3K/Akt signaling up-regulation is potentially more activated in MDAH-2774 than TOV-112D.

It has also been shown that estrogen signaling was found to have a stronger connection with TOV-112D than MDAH-2774 from our IPA analysis. The estrogen receptor (ER) was found to be over-expressed in most ovarian cancers and anti-estrogen drugs have been used to inhibit the growth of ER positive epithelia ovarian cancer cells, implying a strong connection between ER signaling and the tumor, but little is known about the detailed mechanism[28]. The stronger connection with ER signaling in TOV-112D is probably due to the activation of K-ras which has been detected in TOV-112D but not in MDAH-2774 according to our analysis. K-ras is known to be present as the wild type in TOV-112D and mutated in MDAH-2774. Active K-ras can activate the ER through Erk-mediated ER phosphorylation and enhance the steady level of ER. Therefore, ER signaling may turn out to be more pronounced in TOV-112D than MDAH-2774. Also, VEGF and chemokine signaling, which are both related to metastasis, were both shown to be more significant in TOV-112D than MDAH-2774. Other pathways such as insulin-like growth factor-1 (IGF-1) signaling, ERK/MAPK signaling, integrin signaling, cAMP-mediated signaling have all been previously reported to be involved in ovarian cancer[28].

2.4.5.2 Quantitative Analysis

In order to gain further insight into the dynamic changes of the cell state between these two cell lines at the molecular level, we have also sought to examine the differentially expressed components of these pathways in depth. For example, the detailed signaling cascade of PI3K/AKT is depicted in Figure 2.8. By incorporating the normalized spectral count results, we have been able to calculate the relative expression level of identified proteins under this pathway in addition to detecting their presence or absence.

The major network, which is comprised of 34 identified differentially expressed proteins and two imported from IPKB, is displayed in Figure 2.9 with a p-value of $<10^{-49}$. The major functions extracted from this network are related to cancer, reproductive system disease, and skeletal and muscular disease. P53 is the hub of this network, implying that the differential expression level of P53 in these two cell lines is one of the major driving forces for their differentiation in tumor growth.

Pathway analyses of the qualitative data and the quantitative data partially coincide with each other by using IPA. Qualitative data represents a group of proteins with enriched difference between MDAH-2774 and TOV-112D, as they are only detectable in either of these two cell lines. Quantitative data consists of proteins that are detected in both of these two cell lines with a fold change larger than two in addition to those detected in only one of these two cell lines. Analysis based on qualitative data alone is simple to handle, meanwhile it is biased as it excludes the information containing dynamic change in protein abundance whereas quantitative analysis is more comprehensive. The combination of qualitative data and quantitative data is based on the assumption that the spectral count of the protein detected in only one sample is assigned to 0 in the other one. However, we have observed a decrease of sensitivity induced by replacing any missing values with zero. After multiple testing corrections, fold-changes of some proteins between MDAH and TOV are decreased and the q-values are increased, which suggests global signal suppression by this replacement method. One of the explanations could be these missing values are not truly zero, simply because we can not detect them by the current technique. This is especially the case for low abundance peptides which could be masked by their co-eluting high abundance peptides. In the future, target proteomics method (e.g. multiple reaction monitoring methods) will be adopted to verify important proteins.

2.4.6 Western Blot Validation

It is becoming increasingly important to validate the proteome profiling results using alternative technologies. In this study, we used one-dimensional western blot analyses to confirm some of the differential expression results inferred by spectral counting. Three proteins were selected from Table 2.1: UCHL1, Stratifin and MARKS. As can be seen from Figure 2.10, the intensities from these three proteins correlate well with the spectral counting results shown in the left panel.

2.5 Conclusions

Proteomic profiles from two ovarian endometrioid derived cell lines with different genetic mutations have been studied using a shotgun proteomic approach. This involved whole lysate digestion by trypsin with extensive fractionation in the first dimension using cIEF based upon a pH-based separation followed by capillary RP-HPLC. On-line analysis was performed using tandem mass spectrometry acquired by a linear ion trap mass spectrometer. A large number of proteins were identified after filtering through the Peptide Prophet/Protein Prophet Trans Proteomic pipeline. Differentially expressed proteins were quantitated using label free methods and studied by Ingenuity Pathway Analysis (IPA) to reveal the association with important biological functions. It was shown that some important signaling pathways may be highly associated with one of the two cell lines. The PI3K/AKT pathway was found to be significantly predominant in MDAH-2774 but not in TOV-112D. The p53 pathway is shown by network analysis to be important in both cell lines but the network in MDAH-2774 is a more compact one centered on p53 while the network for TOV-112D is more scattered and composed of small networks with ATM, Jnk and GLI1 in the center. The fact that p53 is an important hub of this network implies that this pathway is a major driving force for differentiation and growth. Other pathways such as estrogen signaling were found to have a stronger connection to TOV-112D than MDAH-2774 and activation of K-ras has been detected in TOV-112D but not in MDAH-2774. Thus, the method described can define the important pathways involved in cancer development and how it may differ between samples. This strategy may be important for biomarker discovery and may lead to development of candidates for drug treatment of disease.

2.6 References

[1] McCluskey, L.L., Dubeau, L., Curr Opin Oncol, 1997. **9**(5): p. 465-70.
[2] Scully, R.E., Young, R.H., Clement, P.B, Atlas of Tumor Pathology, 1998, Third Series, Fascicle 23.
[3] Freedman, R. S., Pihl, E., Kusyk, C., Gallager, H. S., and Rutledge, F. 1978, *Cancer* **42,** 2352–2359
[4] Provencher, D. M., Lounis, H., Champoux, L., Tetrault, M., et al, 2000, *In Vitro Cell Dev. Biol. Anim.* **36,** 357–361.
[5] **Wu, R., Zhai, Y., Fearon, E.R., Cho, K.R. et al, *Cancer Research 2001,* 61, 8247-8255**
[6] unpublished data from R.W and K.R.C
[7] Wang, J.H., Hewick, R. M., et al, *Drug Discovery Today*, 1999, 4, 129-133
[8] Wang, H., Kachman, M.T., Schwartz, D.R., Cho, K.R., Lubman, D.M. Proteomics 2004, 4(8), 2476-2495.
[9] Gagne, J.P., Gagne, P., Hunter, J.M., Bonicalzi, M.E., et al, Mol. Cell. Biochemistry, 2005, 275, 25-55.
[10] Washburn, M.P., Wolters, D., Yates, J.R., Nature Biotechnology, 2001, 19 (3): 242-247.
[11] Wang, Y, Balgley, B.M., Rudnick, P.A., Evans, E.L., DeVoe, D.L., Lee C.S., J. Proteome Res , 2005. 4, 36-42.
[12] Zhou, F., Johnston, M.V., Electrophoresis 2005, 26, 1383-1388.
[13] Andreev, V.P., Li, L., Cao, L. Gu, Y., Rejtar, T., Wu, S., Karger, B.L., J. Proteome Research, 2007, 6, 2186-2194.
[14] Wong, J.W.H., Sullivan, M.J., Cagney, G., Briefings in BioInformatics, 2008, 9, 156-165.
[15] Liu, H., Sadygov, R.G., Yates, J.R., Anal. Chem, 2004, 76, 4193-4201.

[16] Silvia M. Uriarte, David W. Powell, Gregory C. Luerman, Michael L. et al, J. Immunol., 2008. 180, 5575 - 5581.
[17] Keller A, Nesvizhskii A.I., Kolker E, Aebersold R., Anal Chem, 2002, 74, 5383-5392.
[18] http://cran.r-project.org/web/packages/qvalue/
[19] Curmi, P.A., Nogues, C., Lachkar, S., Carelle, N., Gonthier, M.P. et al., *Br. J. Cancer* 2000, *82*, 142-150.
[20] Nishio, K., Nakamura, T., Koh, Y., Kanzawa, F., Tamura, T. et al., *Cancer* 2001, *91*, 1494-1499.
[21] Chen, Z., *Cancer* 2001, *92*, 3068-3075.
[22] Jacobberger, J.W., Sramkoski, R.M., Zhang, D.S., Zumstein, L.A., Doerksen, L.D. et al., *Cytometry*, 1999, *38(5)*, 201-213.
[23] Conroy S.E., Latchman D.S. et al, *Br J Cance*r 1996, 74, 717–721
[24] Tonin P.N., Hudson T.J., Rodier F., Bossolasco M, Lee P.D., Novak J, Manderson E.N., Provencher D, Mes-Masson AM. Et al., *Oncogene* 2001, 20, 6617–6626,
[25] Jones M.B., Krutzsch H, Shu H, Zhao Y, Liotta L.A., Kohn E.C., Petricoin, E.F., et al., *Proteomics* 2002, 2: 76–84
[26] Alaiya, A.A., Franzen, B., Fujioka, K., Moberger, B., Schedvins, K. et al., *Int. J. Cancer* 1997, *73*, 678-683.
[27] Campbell, I.G., Russell, S.E., Choong, D.Y., Montgomery, K.G., Ciavarella, M.L., Hooi, C.S., Cristiano, B.E., Pearson, R.B., and Phillips, W.A., Cancer Res 2004, *64*, 7678-7681.
[28] Nicosia, S.V., Bai, W., Cengu, J.Q., Coppola, D., Kurk, P.A., et al., *Hematol Oncol Clin N Am*, 2003, 17, 927-943

Table 2.1: Top10 Expression molecules and top10 differentially expression molecules. The expression values for top10 expression proteins in MDAH-2774 and TOV-112D are normalized mean spectral counts across different runs and are shown in log$_2$ scale. The expression values for top10 differentially expressed proteins are normalized spectral count ratio in log$_2$ scale. Q-value indicates the significance of difference from multiple test correction.

	Protein	Gene	Exp.Valu	qv-value	Description
	Accession	Name	e		
	Number		(log)		
Top10 Expression Molecules in MDAH-2774	P60709	ACTB	8.1		actin, beta
	P08670	VIM	8.04		vimentin
	P38646	HSPA9	7.86		heat shock 70kDaprotein9 (mortalin)
	P68032	ACTC1	7.84		actin,alpha, cardiac muscle1
	P11142	HSPA8	7.75		heat shock 70kDa protein8
	P62736	ACTA2	7.74		actin, alpha2, smooth muscle,
	P10809	HSPD1	7.67		heat shock 60KDa protein1 (chaperonin)
	P043350	TUBB4	7.44		tublin, beta4
	P11021	HSPA5	7.43		heat shock 70KDa protein5 (glucose-regulated protein,78kD)
	P68104	EEF1A1	7.39		eukaryotic translation elongation factor 1 alpha 1)
Top10 Expression Molecules in TOV-112D	P08670	VIM	10.67		vimentin
	P07737	PFN1	8.52		profilin1
	P15531	NME1	8.3		non-metastatic cells 1
	P60709	ACTB	8.2		actin, beta
	P22392	NME2	8.15		non-metastatic cells 2
	P11142	HSPA8	8.03		heat shock 70kDa protein8
	P68104	EEF1A1	7.9		eukaryotic translation elongation factor 1 alpha 1)
	P04083	ANXA1	7.81		annexin A1
	P02461	COL3A1	7.78		collagen, type3, alpha1
	P61978	HNRPK	7.65		heterogeneous nulear ribonucleoproteinK

Table 2.2: Molecules in the top1 network of MDAH-2774 and TOV-112D

Analysis	Molecules in the network	Score	Top Functions
MDAH-2774	BANF1,CIRH1A,Ck2,CKAP2 (includes EG:26586),CSNK2A2, CXXC1,DNA-directedRNApolymerase,F11R,FAM3C,GNL3, HIST1H1C,HIST1H1D,MTDH,PDRG1,PLXNB2,POLR1C,POLR2B, POLR3F,POLR3G,PRIM2,RBBP5,S100A16,SAE1,SARS,SUB1, TCOF1 (includesEG:6949),TEP1,TMED7,Top2,TP53, TUBB4,UBE1L2,UBE2I,UBTF,UPP1	-49	Cell Cycle, Cellular Assembly And organization, DNA Replication, Recombination and Repair
TOV-112D	14-3-3,AOF2,ATM,BNC1,Calcineurin protein(s), CD72,CDC25A,CTBP1,DCTN1,DCX,GATA5(includesEG:140628), GLI1,GLI2,GMFG,GSTM2,H2AFX,HDAC4,HMGA2,Jnk,KRT15, MAP3K3,MAP3K5,MRE11A,NEK2,PKD1,PTHR1,REM1,RFC2, RFC4,RFC5,RFXANK,RPA1,SKI,TP53BP1,ZEB2	-45	Cancer, Cell Cycle, DNA Replication, Recombination and Repair

Figure 2.1a: Experimental Flowchart

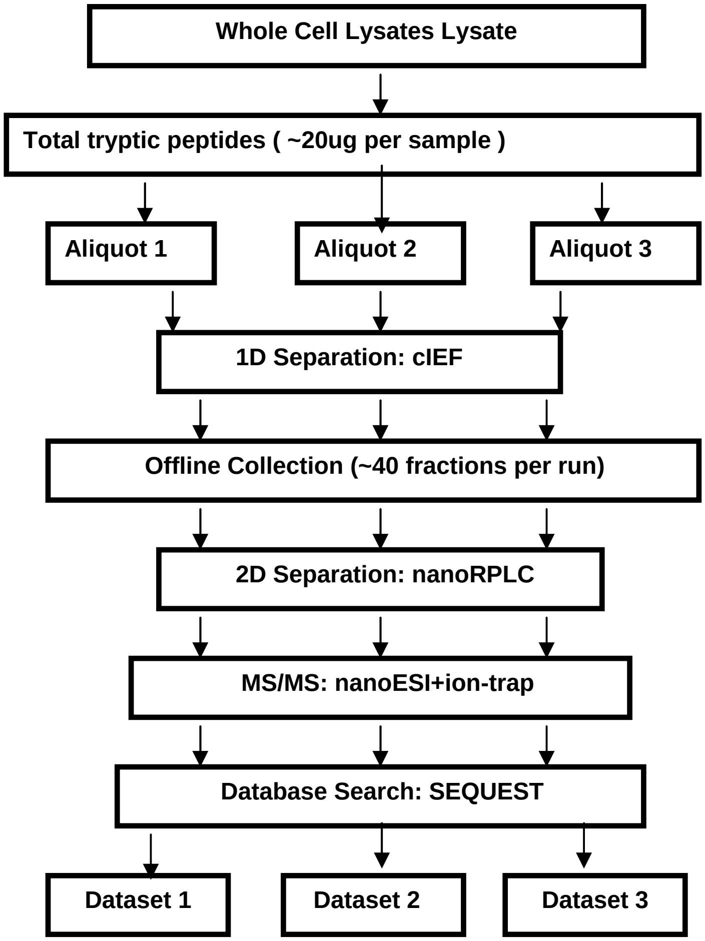


Figure 2.1b: Data Processing Strategy

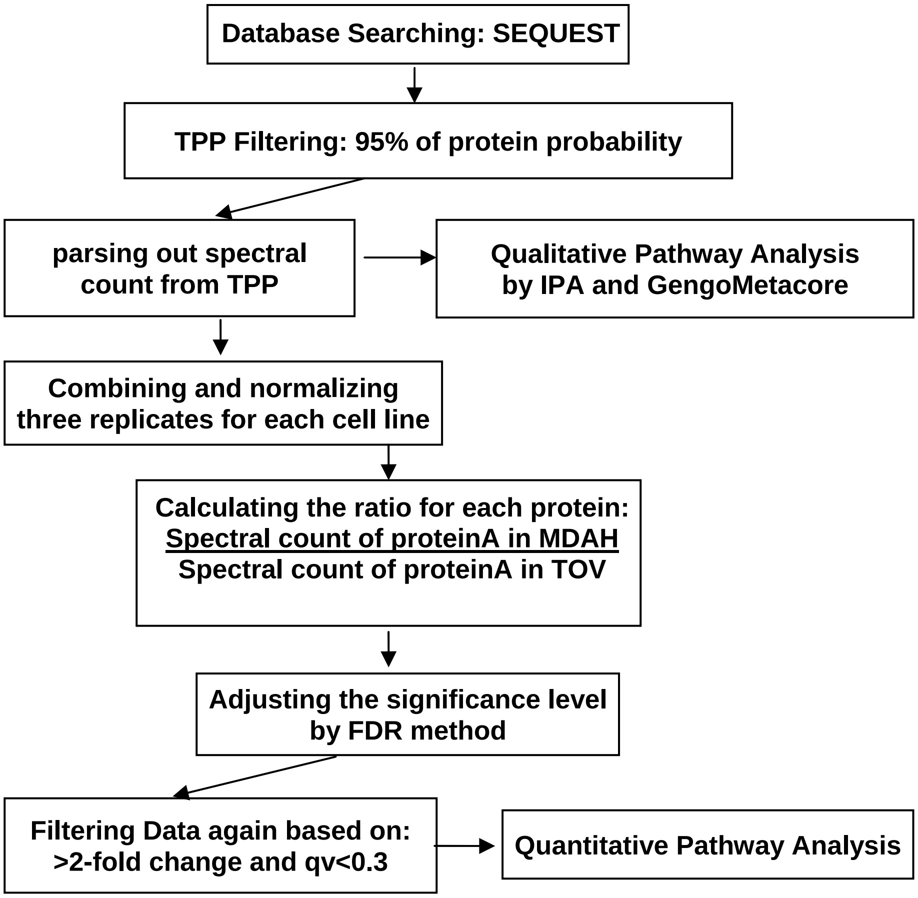

Database Searching: SEQUEST
TPP Filtering: 95% of protein probability
parsing out spectral count from TPP
Qualitative Pathway Analysis by IPA and GengoMetacore
Combining and normalizing three replicates for each cell line
Calculating the ratio for each protein:
Spectral count of proteinA in MDAH
Spectral count of proteinA in TOV
Adjusting the significance level by FDR method
Filtering Data again based on: >2-fold change and qv<0.3
Quantitative Pathway Analysis

Figure 2.2 cIEF-autocollection instrument layout

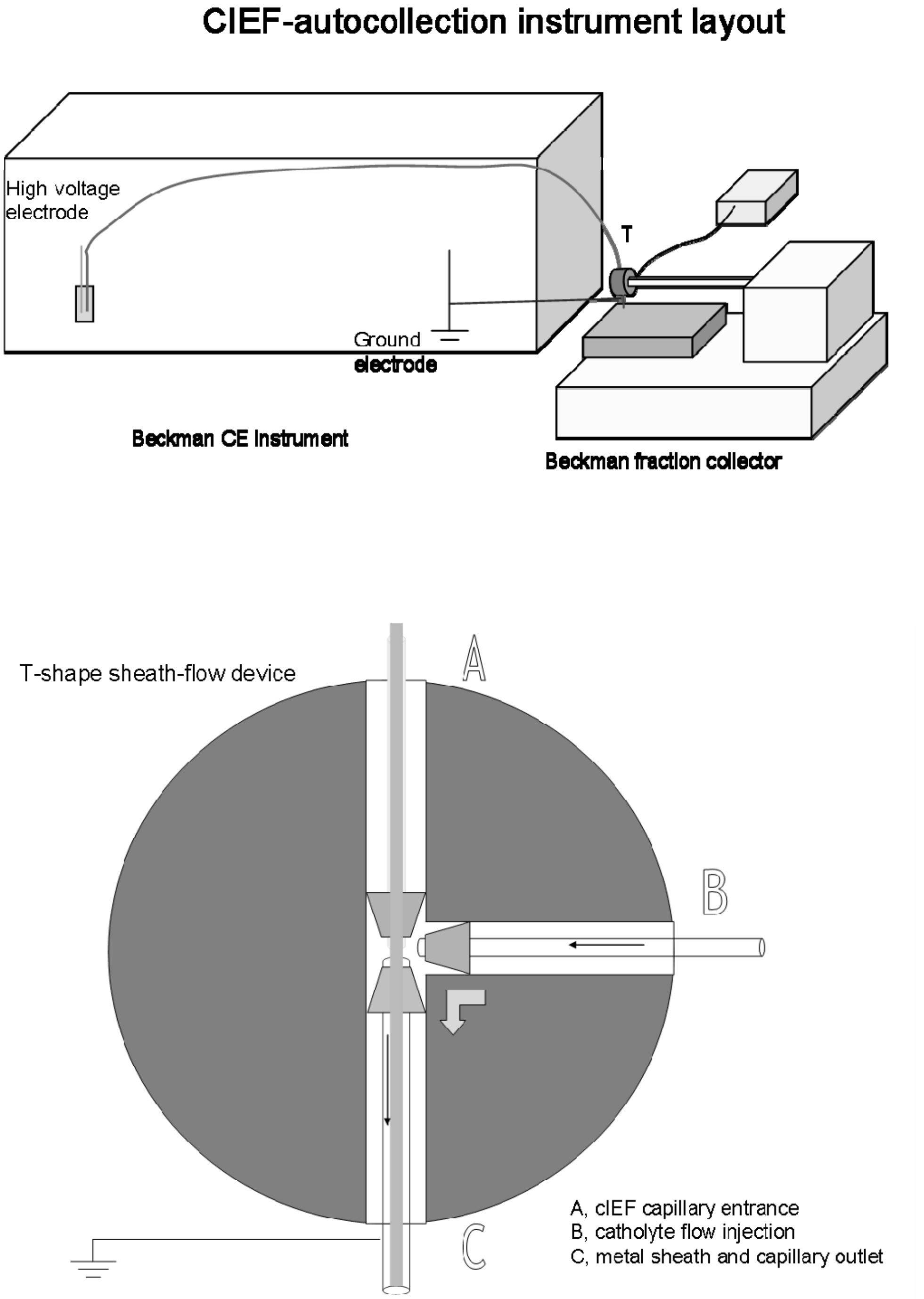

CIEF-autocollection instrument layout
High voltage electrode
T
Ground electrode
Beckman CE instrument
Beckman fraction collector
T-shape sheath-flow device
A
B
C
A, cIEF capillary entrance
B, catholyte flow injection
C, metal sheath and capillary outlet

Figure 2.3: Theoretical pI distribution plot of the second run of MDAH-2774. Fraction number shown in the X-axis is plotted against the average of peptides pI value within each fraction shown in the Y-axis.

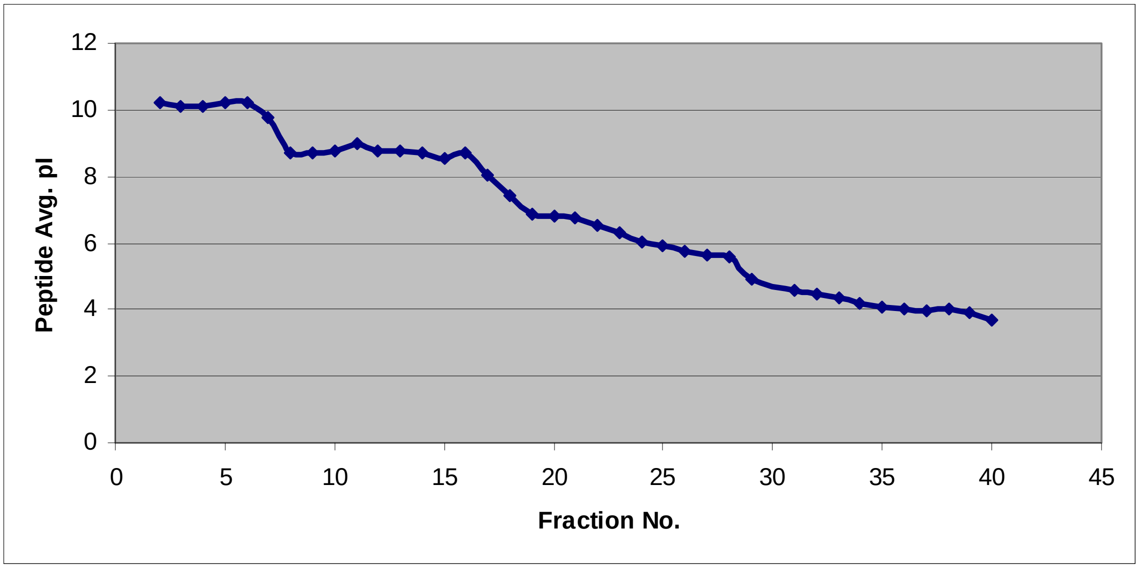
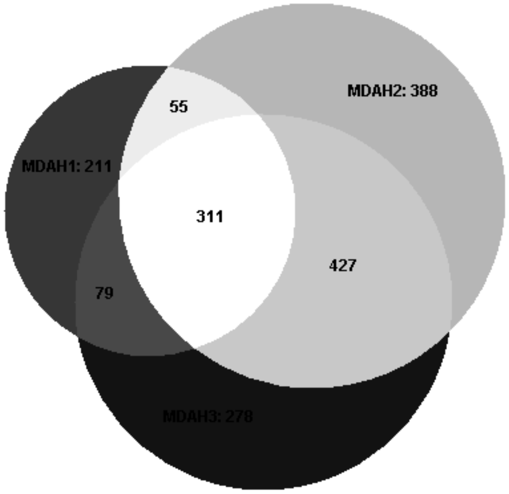
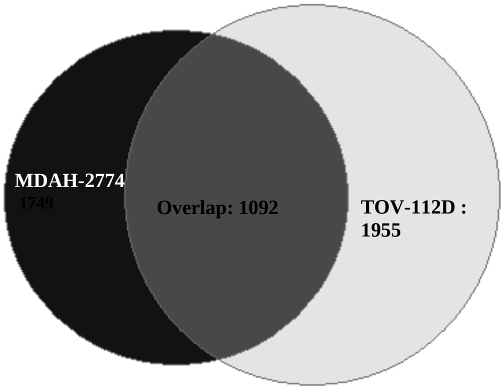


Figure 2.4: Venn diagram of the number of proteins identified from: all three runs of MDAH-2774-2774(2.4a); MDAH-2774-2774 and TOV-112D-112D (2.4b) with a minimum protein probability of 0.95 as given by ProteinProphet^{TM}.

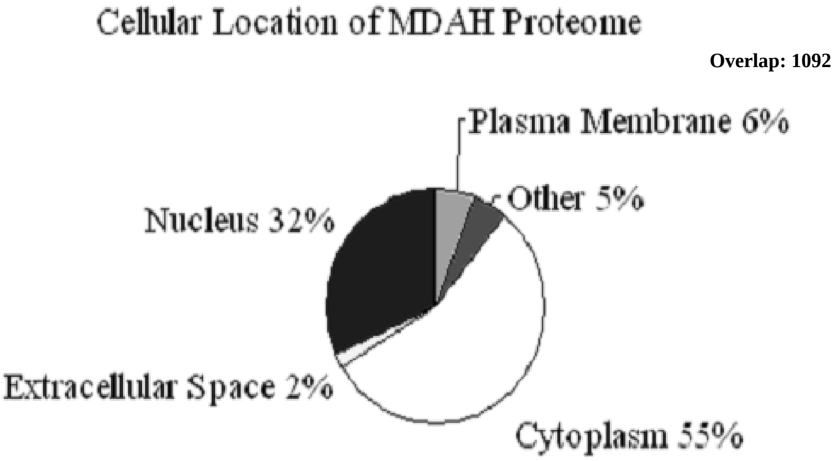


2.4a

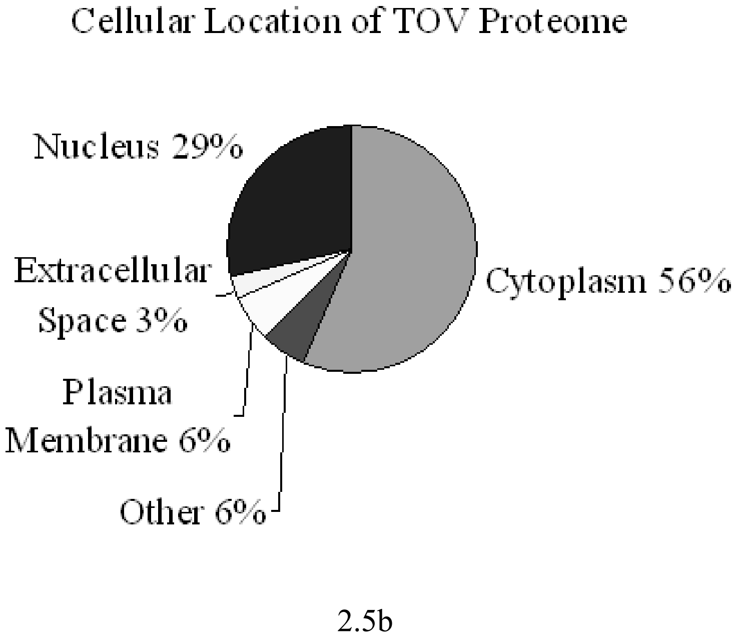


2.4b

Figure 2.5: Cellular Distribution of identified proteins from MDAH-2774 and TOV-112D



2.5a



2.5b

Figure 2.6: Distribution of the protein abundance ratio between MDAH-2774 and TOV-112D on $\log_2$ scale. 828 differentially expressed proteins with fold changes larger than 2 based on normalized spectral count data were used to generate this histogram. Horizontal axis shows the ratio of the relative abundance between filtered proteins from MDAH and TOV on $\log_2$ scale. Vertical axis shows the number of proteins within each column.

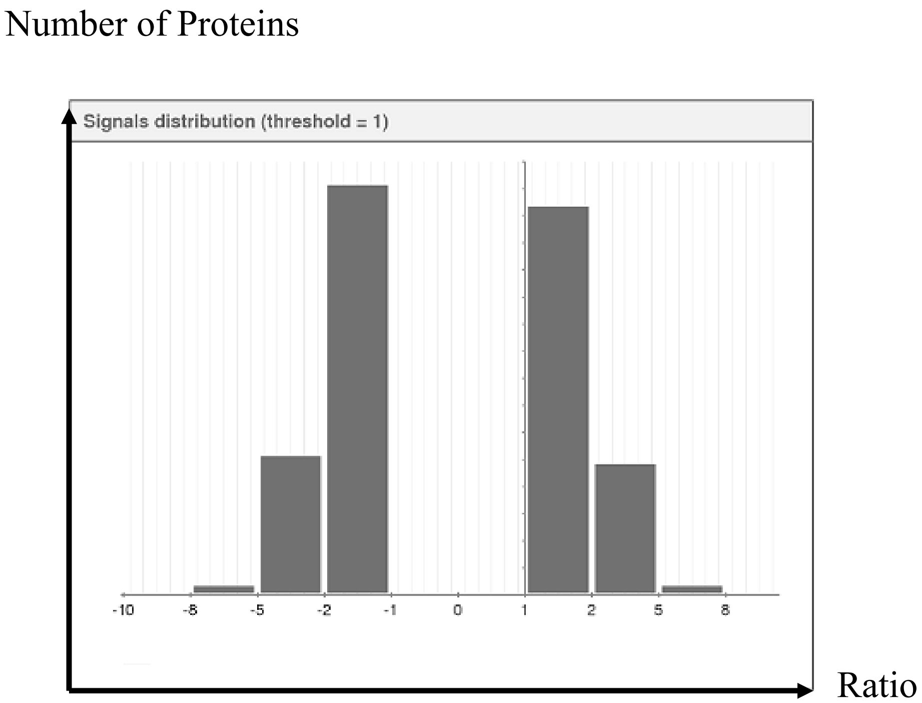


Figure 2.7: Comparison of canonical signaling pathways between MDAH-2774 and TOV-112D. Only the 14 most different pathways are shown in the Figure, as ranked by the significance in MDAH-2774. The vertical line indicates a threshold of p<0.1.

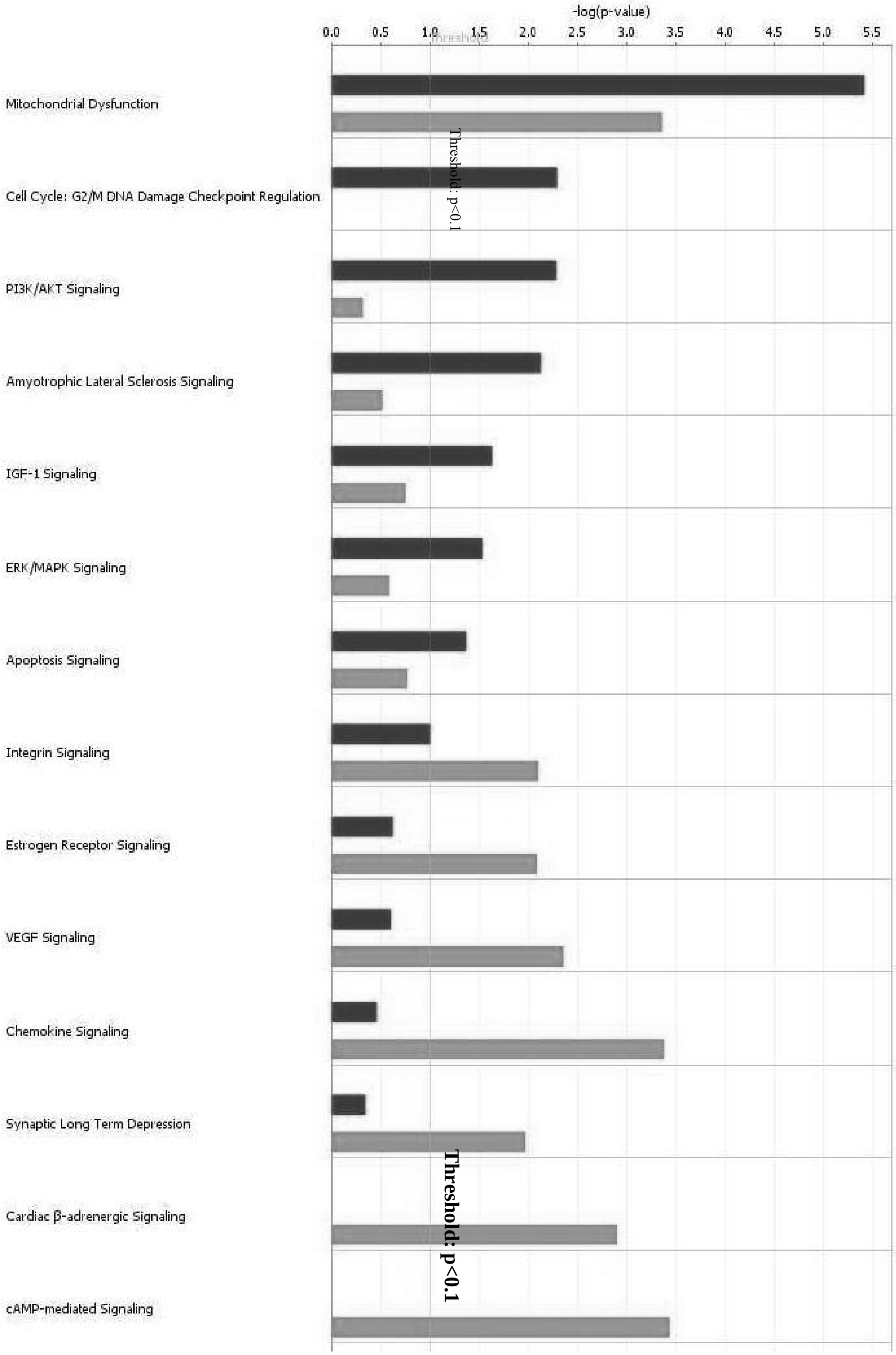


Figure 2.8: Signaling cascade of PI3K/AKT pathway. Green nodes represent over-expression in MDAH-2774 and red nodes represent over-expression in TOV-112D. Plain nodes are imported from IPKB. This figure has been manually modified from integrative analysis by adding some proteins which were identified from only one cell line but did not meet the threshold of integrative analysis.

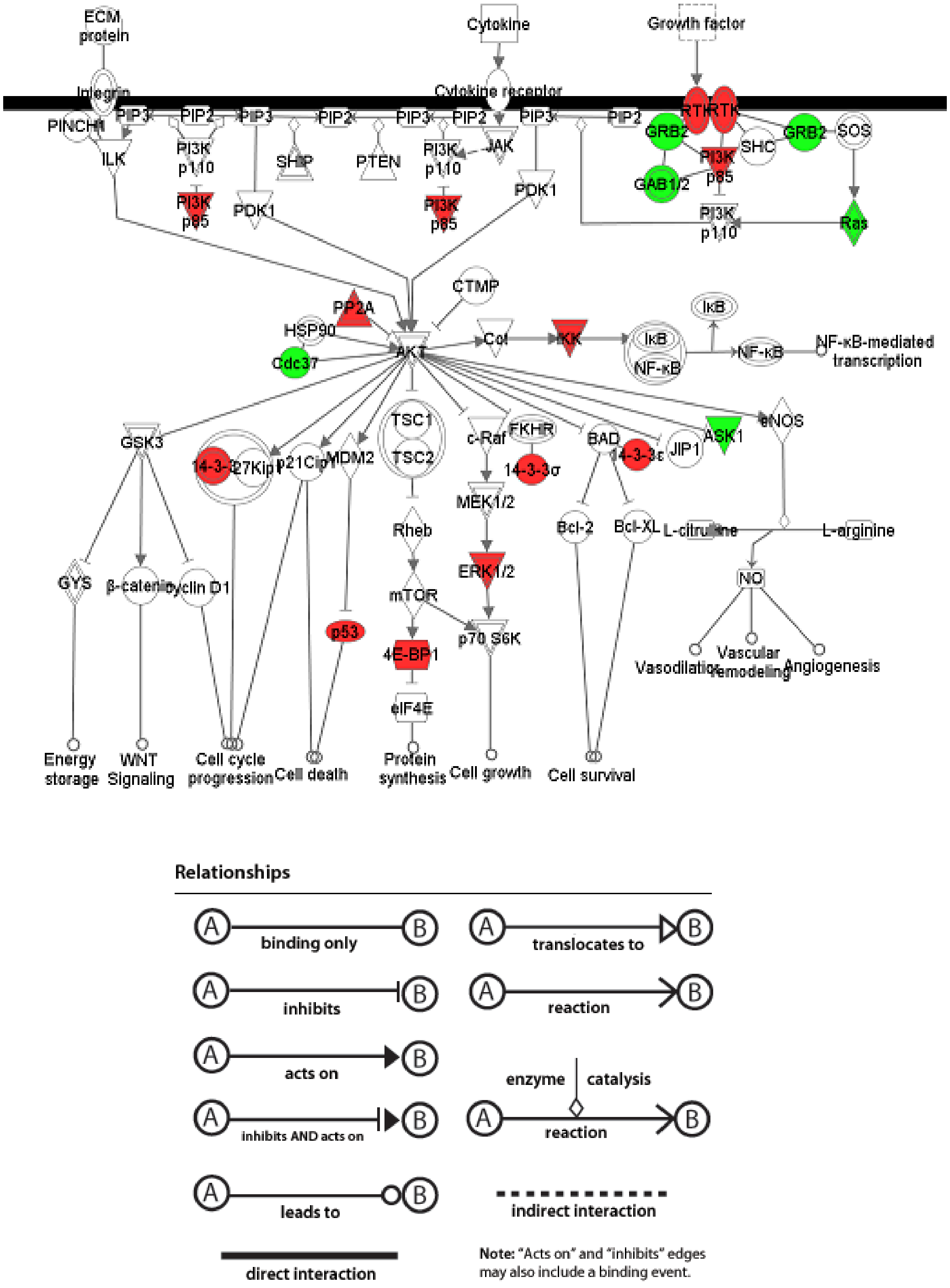


Figure 2.9: Top connectivity network from integrative analysis. Red and green nodes represent proteins that are identified to be over-expressed in MDAH-2774 and TOV-112D respectively. Darker color indicates larger fold-change. The detailed description of these molecules and their relative expression values can be found from supplemental material.

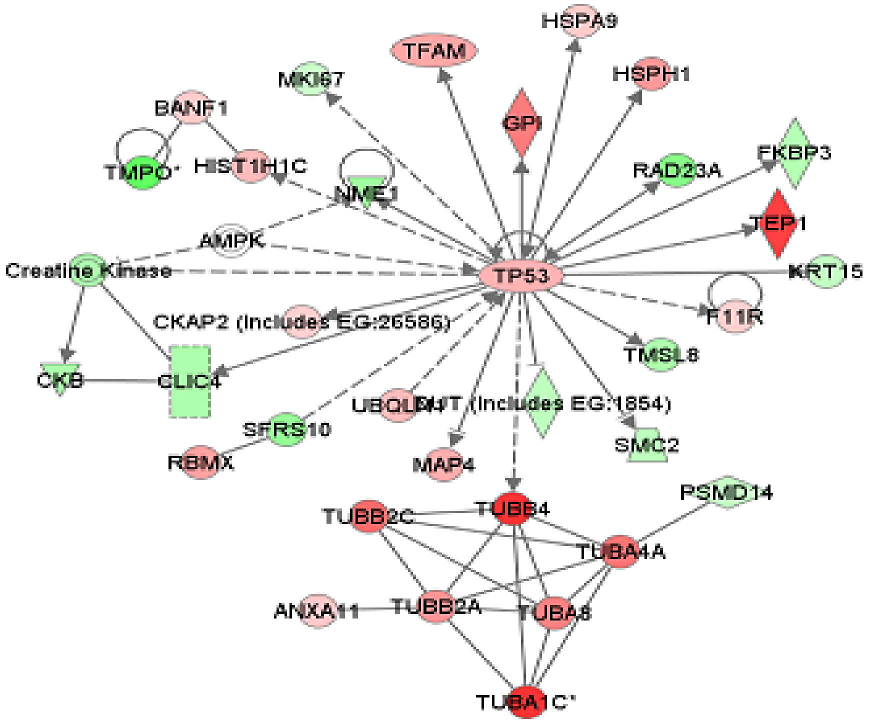


Figure 2.10: Western Results of UCHL1(a), Stratifin(1433-sigma)(b) and MARKS(c). Expression values are normalized spectral count ratio in log2 scale. Positive value indicates over-expression in MDAH-2774 and negative value indicates over-expression in TOV-112D. Band shown on 37KD is Beta-actin which was used as a control.

Figure 2.10a: Western Results of UCHL1

Protein Accession Number	Gene Name	Exp.Value(log)	qv-value
P09936	UCHL1	-6.77	0

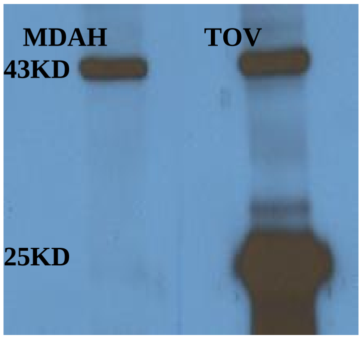


Figure 2.10b: Western Results of Stratifin

Protein Accession Number	Gene Name	Exp.Value(log)	qv-value
P31947	SFN	6.56	0

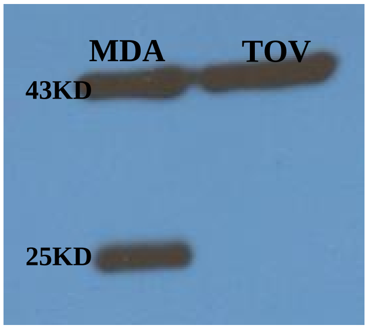
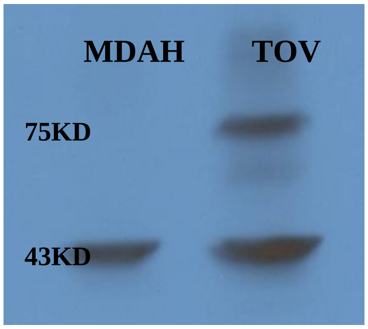


Figure 2.10c: Western Results of MARKS

Protein Accession Number	Gene Name	Exp.Value(log)	qv-value
P29966	MARKS	-5.24	0.01



CHAPTER 3

QUANTITATIVE PROTEOMIC PROFILING STUDIES OF PANCREATIC CANCER STEM CELLS

3.1 Abstract

Analyzing subpopulations of tumor cells in tissue is a challenging subject in proteomic studies. Pancreatic cancer stem cells (CSCs) are such a group of cells that only constitute 0.2-0.8% of the total tumor cells but have been found to be the origin of pancreatic cancer carcinogenesis and metastasis. Global proteome profiling of pancreatic CSCs from xenograft tumors in mice is a promising way to unveil the molecular machinery underlying the signaling pathways. However, the extremely low availability of pancreatic tissue CSCs (around 10,000 cells per xenograft tumor or patient sample) has limited the utilization of currently standard proteomic approaches which do not work effectively with such a small amount of material. Herein, we describe the profiling of the proteome of pancreatic CSCs using a capillary scale shotgun technique by coupling offline capillary isoelectric focusing(cIEF) with nano reversed phase liquid chromatography(RPLC) followed by spectral counting peptide quantification. A whole cell lysate from 10,000 cells which corresponds to ~1ug protein material is equally divided for three repeated cIEF separations where around 300ng peptide material is used in each run. In comparison with a non-tumorigenic tumor cell sample, among 1159 distinct proteins identified with FDR less than 0.2%, 169 differentially expressed proteins are identified after multiple testing corrections where 24% of the proteins are upregulated

in the CSCs group. Ingenuity Pathway analysis of these differential expression signatures further suggests significant involvement of signaling pathways related to apoptosis, cell proliferation, inflammation and metastasis.

3.2 Introduction

Pancreatic cancer has the worst prognosis of any major malignancy and is currently ranked the fourth leading cause of cancer-related mortality with a five-year survival rate less than 5%. Delayed diagnosis, relative chemotherapy and radiation resistance and an intrinsic biological aggressiveness all contribute to the abysmal prognosis[1]. Attempts to better understand the molecular characteristics of pancreatic adenocarcinoma have focused on studying the gene and protein expression profiling of pancreatic cancer compared to either normal pancreas or pancreatitis. However, these studies have not accounted for the heterogeneity of the tumor cells, in particular, the existence of a small set of distinct cells termed cancer stem cells which are responsible for tumor initiation and propagation.

Cancer stem cells have been identified in pancreatic cancer and several other tumor types including colon, prostate, and brain. In Li *et al* [2], a subpopulation of pancreatic tumor cells with cell surface markers $CD44^{+}CD24^{+}ESA^{+}$ was isolated and functional studies were conducted to verify that this subpopulation possessed the ability of self-renewal and producing differentiated progeny. New strategies addressing this disease with a paradigm shift in the mechanism of the therapeutic resistance and recurrence of pancreatic tumor can be developed with an improved understanding of the cellular signaling pathways in CSCs at the protein level. The xenograft model of primary human pancreatic cancer represents a significant advance for the study of pancreatic cancer. Animal models using cancer cell lines often do not recapitulate human diseases accurately, where the biological characteristics and histology of tumor-derived human pancreatic cancer tissue are preserved in the xenograft.

The major obstacle to the study of global proteome expression profiling of the CSCs is the extremely small number of cancer stem cells available per study per tumor sample. Due to the unique features of pancreatic tissue and low percentage of pancreatic cancer stem cells (0.2% to 0.8%), from a single human tumor xenografted in a mouse, we can typically obtain 10k antibody labeled cancer stem cells using flow cytometry, which corresponds to around 1ug of total protein. Current publications of proteomic studies using capillary scale shotgun approaches are based on the analysis of entire tissue sections or cell lines instead of a subpopulation of primary human cells[3]. The amount of material consumed in a proteomic study using shotgun approaches such as MudPIT or offline 2D-LC/MS/MS is higher than 20ug[4-6]. Some studies targeting a certain subpopulation of cells have restricted the analysis to one dimensional separation before mass spectrometry[7, 8] to avoid the sample loss in a 2nd dimension of separation, though limiting the ability of identifying proteins present in lower abundance.

In the present study, we employed a PPS facilitated lysis procedure combined with a high resolution two-dimensional separation to accommodate the small sample size of this study. After cell lysis, protein extracts were digested and equally divided into three aliquots. Each aliquot of around 300ng total material was then introduced into a two-dimensional separation by cIEF and nano-RPLC followed by tandem mass spectrometry analysis to identify the proteins present. A label-free protein quantification using spectral counts was employed to measure the protein fold changes between the CSCs group and the bulk tumor group. Ultimately, the signature proteins detected by the method were then uploaded to Ingenuity Pathway Analysis (IPA) for functional analysis to identify signaling pathways and protein-protein interaction networks that were significant in CSCs compared to bulk tumor cells.

3.3 Materials and Methods

3.3.1 Starting Material Preparation

a. Primary tumor specimen implantation.

Samples of human pancreatic adenocarcinomas were obtained within 30 min following surgical resection according to Institutional Review Board–approved guidelines. Tumors were suspended in sterile RPMI 1640 and mechanically dissociated using scissors and then minced with a sterile scalpel blade over ice to yield two 2–mm pieces. The tumor pieces were washed with serum-free PBS before implantation. Eight-week-old male NOD/SCID mice were anesthetized using an i.p. injection of 100 mg/kg ketamine and 5 mg/kg xylazine. A 5-mm incision was then made in the skin overlying the mid-abdomen, and three pieces of tumor were implanted subcutaneously. The skin incision was closed with absorbable suture. The mice were monitored weekly for tumor growth for 16 weeks.

b. Preparation of single-cell suspensions of tumor cells.

Before digestion with collagenase, low passage primary human pancreatic xenograft tumors from multiple mice were cut up into small pieces with scissors and then minced completely using sterile scalpel blades. To obtain single-cell suspensions, the resultant minced tumor pieces were mixed with ultrapure collagenase IV (Worthington

Biochemicals, Freehold, NJ) in medium 199 (200 units of collagenase per ml) and allowed to incubate at 37 C for 1.5 to 2 hrs for enzymatic dissociation. The specimens were further mechanically dissociated every 15 to 20 min by pipetting with a 10-ml pipette. At the end of the incubation, cells were filtered through a 40-Am nylon mesh and washed with HBSS/20% fetal bovine serum (FBS) and then washed twice with HBSS.

c. Flow cytometry.

Dissociated cells were counted and transferred to a 5-mL tube, washed twice with HBSS containing 2% heat-inactivated FBS, and resuspended in HBSS with 2% FBS at a concentration of 10^6 per 100 ul. Sandoglobin solution (1 mg/ml) was then added to the sample at a dilution of 1:20 and the sample was incubated on ice for 20 min. The sample was then washed twice with HBSS/2% FBS and resuspended in HBSS/2% FBS. Antibodies were added and incubated for 20 min on ice, and the sample was washed twice with HBSS/2% FBS. When needed, a secondary antibody was added by resuspending the cells in HBSS/2%FBS followed by a 20-min incubation. After another washing, cells were resuspended in HBSS/2% FBS containing 4,6-diamidino-2-phenylindole (DAPI; 1 Ag/mL final concentration). The antibodies used were anti-CD44 allophycocyanin, anti-CD24 (phycoerythrin), and anti-H2K (PharMingen, Franklin Lakes, NJ) as well as anti–ESA-FITC (Biomeda, Foster City, CA), each at a dilution of 1:40. In all experiments using human xenograft tissue, infiltrating mouse cells were eliminated by discarding H2K (mouse histocompatibility class I) cells during flow cytometry. Dead cells were eliminated by using the viability dye DAPI. Flow cytometry was done using a FACSAria (BD Immunocytometry Systems, Franklin Lakes, NJ). Side scatter and forward scatter profiles were used to eliminate cell doublets. Cells were reanalyzed for purity, which typically was >97%.

d. Material

10k Pancreatic cancer stem cells and 100k bulk tumor cells were obtained from mice xenografts after sorting by flow cytometry and gently washed three times with cold PBS(pH 7.4) by repetitive pipetting, followed each time by centrifugation at 1000g for 5min at 4°C. In the third time of washing, excessive PBS was gently sucked off with extra caution when cell pellets were observed at the bottom of the tube.

3.3.2 Cell lysis and Trypsin digestion

PPS (Protein Discovery, Knoxville, TN) powder was dissolved in 50 mM Ammonia Bicarbonate and was added to each tube at a final concentration of 0.2%(m/v). Around a 100 ul cell suspension was then vortexed and incubated at 60°C for 10min, followed by sonication in an ice-water bath for 2 hrs. An aliquot of 5 mM DTT was added and the mixture was incubated at 60°C for 30min. After cooling, 15 mM iodoacetamide was added and the mixture was placed in the dark at room temperature for 30 min in order to allow the carboxymethylation reaction of cysteine residues. 50 mM ammonia bicarbonate was then added at a dilution ratio of 1:5 and 1:50(w/v) L-1-tosylamido-2-phenyletyl chloromethylketone modified sequencing-grade porcine trypsin (Promega, Madison, WI) was added. The mixture was incubated at 37°C in a water bath with agitation. Formic acid (FA) was then added to make a final concentration of 2% to stop the proteolysis. Following termination, the acidified mixture was placed in a 37°C water bath again for 4 hrs to facilitate the hydrolysis and allow the cleavage of PPS. The acidified tryptic peptide mixture was then desalted by a peptide micro-trap (Michrom, Auburn, CA) and eluted with 98% acetonitrile (ACN) and 0.3% FA, followed by spinning to dryness using a SpeedVac concentrator (Labconco, Kansas City) and stored in the -80°C freezer for future use. All chemicals were purchased from Sigma unless mentioned otherwise.

3.3.3 First Dimensional Separation: cIEF

A Beckman CE instrument was modified for cIEF with fraction collection. A 80 cm cIEF (100 mm id, 365 mm od) capillary was coated as previously described[10]. Lyophilized peptides were first reconstituted in gel buffer containing 2% ampholyte (pH 3-10) and were injected hydrodynamically to fill the capillary. Peptide focusing was performed by applying 21kv voltage to the two ends of the capillary using 0.1M phosphoric acid and 1mM sodium hydroxide as the anolyte and catholyte, respectively. The cathodic end of the capillary was kept in a stainless steel coaxial device. As the current reached its plateau, the focusing was complete and the focused peptides were mobilized under pressure and eluted into a 96-well auto-sampler plate by a 2uL/min flow of catholyte solution delivered by a syringe pump. The auto-sampler plate was moved from well to well automatically every 2 minutes by a Beckman fraction collector.

3.3.4 Second dimensional separation: nanoRPLC+nanoESI-MS/MS

When cIEF separation was completed, all cIEF runs with each containing ~30 pI fractions were injected in randomized order via Paradigm auto-sampler (Michrom Biosciences, Auburn, CA) and loaded onto a desalting nano-trap (300 x 50mm) (Michrom) connected to a nano-RP column (C18AQ, 5µm 200A, 100 x 150 mm)(Michrom) by a Paradigm AS1 micro-pump (Michrom). The mobile phases A and B were composed of 0.3% FA in water and 0.3% FA in ACN, respectively. Peptides were first desalted and enriched starting at 100%A with a flow rate of 50 µl/min for 5 min. The sample was subsequently separated by a nano-RP column with a flow rate of 0.3 µl/min after splitting. The linear gradient for separation was as follows: from 3% ACN to 12% ACN in 5 min, from 12% ACN to 40% ACN in 30min, from 40% ACN to 80% ACN in 15 min and finally decreased from 80% ACN to 3% ACN in 10min. The resolved peptides were then introduced into a ThermoFinnigan linear ion trap mass spectrometer (LTQ) (Thermo Electron, San Jose, CA) equipped with a nano-spray ion source (Thermo Electron). The LTQ was operated in data dependent mode in which one cycle of experiments consisting of one full MS scan was followed by five pairs of zoom scans and MS/MS scans with dynamic exclusion set to 30 s. The capillary temperature was set at 175°C, spray voltage was 2.8 kV, capillary voltage was 30 V and the normalized collision energy was 35% for the fragmentation.

3.3.5 Database Search and Protein Identification

MS/MS spectra were then searched against the human UniProt FASTA database by TurboSEQUEST provided by Bioworks ver3.1 SR1 (Thermo-Finnigan). The following modification was allowed in the search: 15.99 Da shift for oxidized Met residues; 58.1 Da shift for carboxymethylated Cys residues. The identified peptides were subsequently processed through PeptideProphet and ProteinProphet incorporated in the trans-proteomic pipeline (TPP: http://proteinprophet.sourceforge.net/prot-software.html) where each protein was assigned with a probability indicating the significance level of the protein appearing in the original sample. In this study, we used a protein probability score of 0.99 as the threshold for protein identification and the FDR is below 0.2%.

3.3.6 Label-free Protein Quantitation and Data Transformation

Spectral counts were parsed out of TPP xml files after processing the SEQUEST data and used as a surrogate measure of protein abundance in our analysis. Global normalization was used to reduce technical bias when acquiring spectral count data from different runs between and across samples. The bias may come from instrument error or the inherent random sampling nature of the LTQ. Three replicated datasets for the CSCs group and four replicated datasets for the tumor group (denoted as CSC1, CSC2, CSC3 and tumor1, tumor2, tumor3, tumor4), containing proteins with 99% confidence or above, together with their spectral counts, were generated. The data were consolidated to form a matrix with seven columns; missing values were replaced with zero. To eliminate the discontinuity observed in simple count ratios when a protein shows spectral count 0, raw data were transformed according to Old *et al*[10], as originally proposed by Beissbarth *et al*[25] for serial analysis of gene expression (SAGE). The transformation uses the $\log_2$ scale quantity

$$N= \log_2[(n+f)/(t-n+f)] \qquad \text{(Eq.1)}$$

for each protein, where n is the raw (globally normalized) spectral count value; t is the total number of spectra over all proteins in each dataset; and f is a correction factor. Larger values of f shrink the results for low spectral count proteins toward zero, thereby eliminating the discontinuity at zero and down weighting the results with greatest measurement error (i.e. the proteins with low spectral counts). Several procedures for setting the constant term f have been proposed; we devised a new approach that is suitable for experiments with technical replicates. We considered the following criterion:

$$R(f) = \Sigma \text{ cor(between replicates)} \qquad \text{(Eq.2)}$$

Correction factor f is defined to be the value that maximizes the correlation given in (Eq. 2). This maximizing effect yielded the value f=3. A schematic view of this computational process is depicted in Figure 3.1(b).

After transformation, statistical significance levels between the CSCs group and the tumor group were then determined by student's t-test followed by multiple testing adjustment using FDR test. Differentially expressed proteins used for subsequent pathway analysis were declared at the level of q-value < 0.1. This FDR test was performed using R-package (http://cran.r-project.org/web/packages/fdrtool/index.html). Fold change (FC) is computed from transformed data using the mean of spectral counts from all replicates within a group: FC = (mean CSCs group) – (mean tumor group), where a positive sign indicates over expression in the CSCs group and a negative sign indicates over expression in the tumor group.

3.3.7 Ingenuity Pathway Analysis (IPA)

To obtain detailed molecular information and infer significant signaling pathways from our global profiling results, differentially expressed proteins from the CSCs group and tumor group were uploaded to IPA. The uploaded Excel spreadsheet file contains the relevant proteins with their fold change, q-value and corresponding primary accession number. The significance values for canonical pathways were calculated using the right-tailed Fisher's Exact Test by comparing the number of proteins that participate in a given function or pathway relative to the total number of occurrences of these proteins in all functional/pathway annotations stored in the ingenuity pathway knowledge base (IPKB).

3.4 Results and Discussion

3.4.1 Evaluation of cIEF+RPLC platform

Capillary isoelectric focusing is a powerful 1st dimensional separation for protein/peptides because of its high resolution and orthogonal separation mechanism versus RP-HPLC. This pI based separation provides an optimal resolution of 0.01 pH unit, which indicates a peak capacity of 700 in a pH range from 3 to 10, while strong cation exchange only has a peak capacity of around 50. SCX also presents undesired retention of peptides with strong interaction with the chromatographic resin and results in poor sample recovery rate. In contrast, cIEF is performed in an open capillary which is usually neutrally coated to prevent electroosmotic flow (EOF) and absorption of samples. Thus cIEF usually provides a sample recovery rate of higher than 90% which is critical in analyzing the extremely small amount of sample in our pancreatic CSCs study.

The quality of the cIEF separation in terms of the resolution and reproducibility is essential to the accuracy of the comparative proteomic study. The theoretical pI value for each identified peptide within each fraction was calculated after database searching. The pI distribution plot from the first replicate of the CSCs group is shown in Figure 3.2(a). As expected, the pIs of the fractions decrease from 10 to 4 following a linear trend except the first 7 fractions where only a few proteins were identified. Peptides beyond the pH range of the ampholytes (pH 3–10) used in these experiments were not expected to be resolved. Overall, cIEF exhibited high separation resolution where more than 70% of the peptides were identified in no more than a single fraction. In addition, the off line collection method has the advantage of maintaining the separated peptide bands, whereas on line collection directly coupled to RPLC reduces the workload at the cost of sacrificing separation resolution and increased sample loss. The number of peptides identified from the four replicate runs of the same tryptic digest from the tumor sample is plotted against their pI values shown in Figure 3.2(b). The distribution of the peptides demonstrates excellent technical reproducibility of the experiment.

Reproducibility is also accessed by pairwise comparison of two selected replicates (CSCs replicate run 1 and 2, Tumor replicate run 1 and 2) using the Pearson correlation coefficient. No transformation was performed at this point. Common proteins identified in both replicate runs were used to calculate the correlation values (R) which are 0.87 with 95% confidence level between [0.84, 0.9] for CSC1 vs CSC2 shown in Figure 3.2(c) and 0.91 for tumor1 vs tumor2 shown in Figure 3.2(d) with 95% confidence level between [0.89, 0.93] shown in Figure 3.2(d).

3.4.2 Spectral Counting Results and Transformation

Detecting relative protein quantity change between various disease classes or different samples is central to understanding the molecular processes of the cell. Currently there are two widely used but fundamentally different label-free protein quantification strategies: spectral counting and peak-area measurement. The latter method requires aligning the retention time of the chromatogram peaks to accurately locate the same peptide and is preferred in experiments where peptides are subject to only one-dimensional separation before being analyzed by mass spectrometry for its accuracy, while the peak-area method is difficult to use with approaches containing two-dimensional separations where the same peptide might appear in more than one LC/MS/MS run. Spectral counting is relatively easy to apply to two-dimensional separation data. We have previously reported a study using spectral counting to analyze the protein expression levels in two ovarian cancer cell line samples and spectral counting results of selected proteins were all consistent with western blot experiments[9].

Different types of computational algorithms regarding the processing strategies of spectral counting data have been proposed by a number of groups[10-12]. Boris et al.[11] proposed a NSAF method to normalize and transform spectral count data based on protein length whereas Old et al.[10] adopted a transformation method to avoid the discontinuity problem which was originally used for serial analysis of gene expression. In the present study, spectral counts were assigned to each identified protein followed by global normalization by adjusting the mean of each dataset to be equal. After consolidating all 7 datasets into a matrix and replacing missing values with zero, we proceeded to identify the correction parameter f that produces the most meaningful abundance measurements for our experimental data. The transformation (Eq. 1) that we applied has the effect of shrinking the expression scores for low spectral count proteins toward zero, compensating for their relatively greater uncertainty. By maximizing R(f) (Eq. 2), the data was transformed in such a way that low-abundance proteins were appropriately, but not excessively down weighted in the analysis. The approach to defining f by maximizing R(f) is based on the fact that technical replicates from the same biological sample should be intrinsically the same. The correction factor was calculated to be 3 and then the whole dataset was transformed accordingly. This is similar to Old's method[10] that f is adjusted according to higher correlation between expected and observed. The fitted transformation, shown in Figure 3.3, indicated that above 300 expression units on the original scale, the measured data were sufficiently reliable to be used without adjustment; between 100 and 300 expression units on the original scale the data were substantially shrunk toward zero, but still contributed to the analysis; and below 100 units, the data were shrunk to the degree that they have little influence in the analysis.

An important issue with any data transformation scheme that uses the class labels or clinical outcome information is the possibility that it may induce artifactual correlations between protein abundance and outcomes. The transformation function (Eq. 1) was monotonic, meaning that if the spectral counts of protein A were larger than those of protein B, this relationship would continue to hold after the transformation is applied. A monotonic transformation is limited in its ability to induce spurious correlations. To further explore whether the transformation induced spurious correlations, we conducted a simulation study. We simulated spectral counts for 1000 proteins measured in two groups of samples, each with three replicates. All 6,000 data values were simulated independently from a standard exponential distribution. We applied the procedure described above for a range of f values from 0 to 100, and considered the number of Z-statistics (comparing the two groups of three samples for a given protein) that exceeded various thresholds (2, 2.5, and 3). We did not observe any inflation in the number of significant associations for f>0 compared to f=0. Furthermore, we observed in our experimental data that the results were not very sensitive to the specific value of f, as long as it was not too close to zero. Figure 3.4(a) shows the clustering results after transforming by f =3. The CSCs group and the tumor group are well separated without introducing artificial interference.

3.4.3 Protein Profiling Results

Accepted protein identifications were obtained after applying a cutoff protein probability of 0.99 by TPP which ensures the FDR is below 0.2%. In addition to this bayes approach based FDR method, we also tested the identification confidence by applying a target decoy database search on one of the tumor datasets. A 0.9 protein probability filter resulted in a FDR of 1.4% and a 0.99 protein probability filter further decreased the decoy hits to zero. A total of 763 and 1031 distinct proteins were identified from three CSCs replicates and four tumor replicates, respectively.

Each identified protein was assigned a cellular location based on information from IPKB. Figure 3.5(a) and Figure 3.5(b) show the cellular distribution of 763 and 1031 identified proteins from the CSCs group and the tumor group, respectively. The cellular distribution is consistent for both of these two groups: the majorities are cytoplasmic and nuclear proteins; plasma membrane proteins occupy 10% of each total proteome, suggesting PPS has the ability to extract hydrophobic proteins. Multiple correcting testing (FDR) was then performed to capture the differentially expressed molecules and 161 out of 1159 proteins were identified by using a threshold of q-value < 0.1. 24% of these differentially expressed proteins show up-regulation in the CSCs group and a few of them are related to the key signaling pathways of CSCs. For example, inter-alpha trypsin inhibitor H3 (ITIH3) was identified in all 3 runs of the CSCs group with two unique peptides and an average spectral count larger than 10, while it was only identified in one of the tumor runs with low spectral counts. This protein associated with inflammatory response in local tissue[13]was previously reported to be one of the downstream target genes of Sonic Hedgehog (Shh), which plays a key role in signaling pathways that directly activate the genes involved in the self-renewal and apoptosis-inhibition functions of CSCs[14] . The over-expression of ITIH3 is consistent with previous finding of the up-regulated Shh at the mRNA level in pancreatic CSCs[2]. In contrast, a mitochondrial apoptosis-inducing factor (AIFM1) was down-regulated in the CSCs group. This protein was identified in all 4 runs of the tumor group with 5 unique peptides, whereas it was only identified in one of the CSCs runs. The decreased protein level detected in the stem cell group from our study agrees with previous reports that inactivation of AIFM1 renders embryonic stem cells resistant to cell death [15]. Beside these significantly differentially expressed proteins, some important low-abundant proteins were also detected although they were not found to be significantly different based on the FDR test. For example, NF-κβ was identified in one of the CSCs runs with two unique peptides but not in any of the tumor runs, as well as c-MET and CXCL5. Their absence in the tumor group is mainly due to their low abundance level which was below the detection limit, however their relative over expression in the CSCs group agrees with previous findings [16-20] that the elevated expressions of these proteins are related to the properties of CSCs.

3.4.4 Signaling Pathway and Connectivity Network Analysis

A list of 169 differentially expressed proteins was uploaded into IPA for functional annotation and pathway analysis. The most variant and relevant canonical signaling pathways enriched with differentially expressed molecules between the CSCs group and the tumor group were generated by IPA and are ranked by significance shown in Figure 3.6 with a threshold of p-value < 0.1. The length of the bars indicate the significance of the signaling pathways to which the differentially expressed proteins are related.

These enriched pathways can be grouped into four main categories related to the characteristics of cancer stem cells: resistance to apoptosis, dysregulation of cell proliferation, association with inflammation and metastasis. The top pathway, Mitochondrial Dysfunction, is related to both apoptosis and tumorigenesis[21, 22] and a recent report has linked mitochondrial dysfunction to ovarian cancer stem cells[23]. Pathways mapped to cellular growth, proliferation and development by IKGB include ILK Signaling, RhoA Signaling and Integrin Signaling. CXCR4 signaling and Acute Phase Response signaling pathways categorized under cellular immune/inflammatory response are also shown to be significantly involved. The connection between inflammation and tumorigenesis has been recognized in many pathologic conditions including pancreatic cancer where some clues are suggesting that inflammation might induce an accelerated process of mutagenesis and mutation accumulation[1, 24]. VEGF signaling associated with angiogenesis is also shown to be significant.

To further infer the functional relevance between these differentially expressed molecules, we have constructed connectivity networks by IPA. The top network which has the highest score of 54 (p-value < $1x10^{-54}$ from Fisher Exact Test) consists of 27 signature proteins shown in Figure 3.7. Interestingly, NF-κβ is imported as the central node to generate this interaction network by IPA, suggesting a potential involvement of this transcription factor although this molecule is not in the experimental signature protein list.

3.5 Concluding Remarks

This work represents the first proteome profiling study on pancreatic cancer stem cells from xenograted tumors in mice. We overcome the difficulty of analyzing the extremely small number of cancer stem cells from the xenografted tumor by using an ultrasensitive sample preparation procedure and cIEF as the 1st dimensional separation before LC/MS/MS to minimize sample loss and obtain high resolution of peptides. A modified transformation algorithm was also devised to handle spectral counting data with technical replicates in order to weight proteins with a large range of spectral counts appropriately in the subsequent statistical analysis. 169 proteins have been captured as differentially expressed signatures between the CSCs group and the bulk tumor group. Pathway analysis and network modeling by IPA has further revealed significantly involved signaling cascades relevant to the characteristics of CSCs.

It will be important to validate proteome profiling results using alternative technologies such as the Western blotting or RT-PCR. However, currently, no other techniques work effectively with less than 1ug of protein material. To compensate for the lack of validation, we increased the confidence level of protein identifications by adopting a more stringent threshold. We set the cutoff to a protein probability score of 0.99 which has a FDR less than 0.2% compared to the commonly used score of 0.9 which gives FDR less than 1%. Moreover, multiple testing corrections are employed to control false positives induced by performing significance tests on a large number of proteins. Thus, differentially expressed proteins identified in this way are more likely to be true positives.

3.6 References

[1] McCluskey, L.L., Dubeau, L., Curr Opin Oncol, 1997. **9**(5): p. 465-70.
[2] Scully, R.E., Young, R.H., Clement, P.B, Atlas of Tumor Pathology, 1998, Third Series, Fascicle 23.
[3] Freedman, R. S., Pihl, E., Kusyk, C., Gallager, H. S., and Rutledge, F. 1978, *Cancer* **42,** 2352–2359

[4] Provencher, D. M., Lounis, H., Champoux, L., Tetrault, M., et al, 2000, *In Vitro Cell Dev. Biol. Anim.* **36,** 357–361.
[5] **Wu, R., Zhai, Y., Fearon, E.R., Cho, K.R. et al, *Cancer Research 2001,* 61, 8247-8255**
[6] unpublished data from R.W and K.R.C
[7] Wang, J.H., Hewick, R. M., et al, *Drug Discovery Today*, 1999, 4, 129-133
[8] Wang, H., Kachman, M.T., Schwartz, D.R., Cho, K.R., Lubman, D.M. Proteomics 2004, 4(8), 2476-2495.
[9] Gagne, J.P., Gagne, P., Hunter, J.M., Bonicalzi, M.E., et al, Mol. Cell. Biochemistry, 2005, 275, 25-55.
[10] Washburn, M.P., Wolters, D., Yates, J.R., Nature Biotechnology, 2001, 19 (3): 242-247.
[11] Wang, Y, Balgley, B.M., Rudnick, P.A., Evans, E.L., DeVoe, D.L., Lee C.S., J. Proteome Res , 2005. 4, 36-42.
[12] Zhou, F., Johnston, M.V., Electrophoresis 2005, 26, 1383-1388.
[13] Andreev, V.P., Li, L., Cao, L. Gu, Y., Rejtar, T., Wu, S., Karger, B.L., J. Proteome Research, 2007, 6, 2186-2194.
[14] Wong, J.W.H., Sullivan, M.J., Cagney, G., Briefings in BioInformatics, 2008, 9, 156-165.
[15] Liu, H., Sadygov, R.G., Yates, J.R., Anal. Chem, 2004, 76, 4193-4201.
[16] Silvia M. Uriarte, David W. Powell, Gregory C. Luerman, Michael L. et al, J. Immunol., 2008. 180, 5575 - 5581.
[17] Keller A, Nesvizhskii A.I., Kolker E, Aebersold R., Anal Chem, 2002, 74, 5383-5392.
[18] http://cran.r-project.org/web/packages/qvalue/
[19] Curmi, P.A., Nogues, C., Lachkar, S., Carelle, N., Gonthier, M.P. et al., *Br. J. Cancer* 2000, *82*, 142-150.
[20] Nishio, K., Nakamura, T., Koh, Y., Kanzawa, F., Tamura, T. et al., *Cancer* 2001, *91*, 1494-1499.
[21] Chen, Z., *Cancer* 2001, *92*, 3068-3075.
[22] Jacobberger, J.W., Sramkoski, R.M., Zhang, D.S., Zumstein, L.A., Doerksen, L.D. et al., *Cytometry*, 1999, *38(5)*, 201-213.
[23] Conroy S.E., Latchman D.S. et al, *Br J Cance*r 1996, 74, 717–721
[24] Tonin P.N., Hudson T.J., Rodier F., Bossolasco M, Lee P.D., Novak J, Manderson E.N., Provencher D, Mes-Masson AM. Et al., *Oncogene* 2001, 20, 6617–6626,
[25] Jones M.B., Krutzsch H, Shu H, Zhao Y, Liotta L.A., Kohn E.C., Petricoin, E.F., et al., *Proteomics* 2002, 2: 76–84
[26] Alaiya, A.A., Franzen, B., Fujioka, K., Moberger, B., Schedvins, K. et al., *Int. J. Cancer* 1997, *73*, 678-683.
[27] Campbell, I.G., Russell, S.E., Choong, D.Y., Montgomery, K.G., Ciavarella, M.L., Hooi, C.S., Cristiano, B.E., Pearson, R.B., and Phillips, W.A., Cancer Res 2004, *64*, 7678-7681.
[28] Nicosia, S.V., Bai, W., Cengu, J.Q., Coppola, D., Kurk, P.A., et al., *Hematol Oncol Clin N Am*, 2003, 17, 927-943

Figure 3.1(a): Experimental flow chart

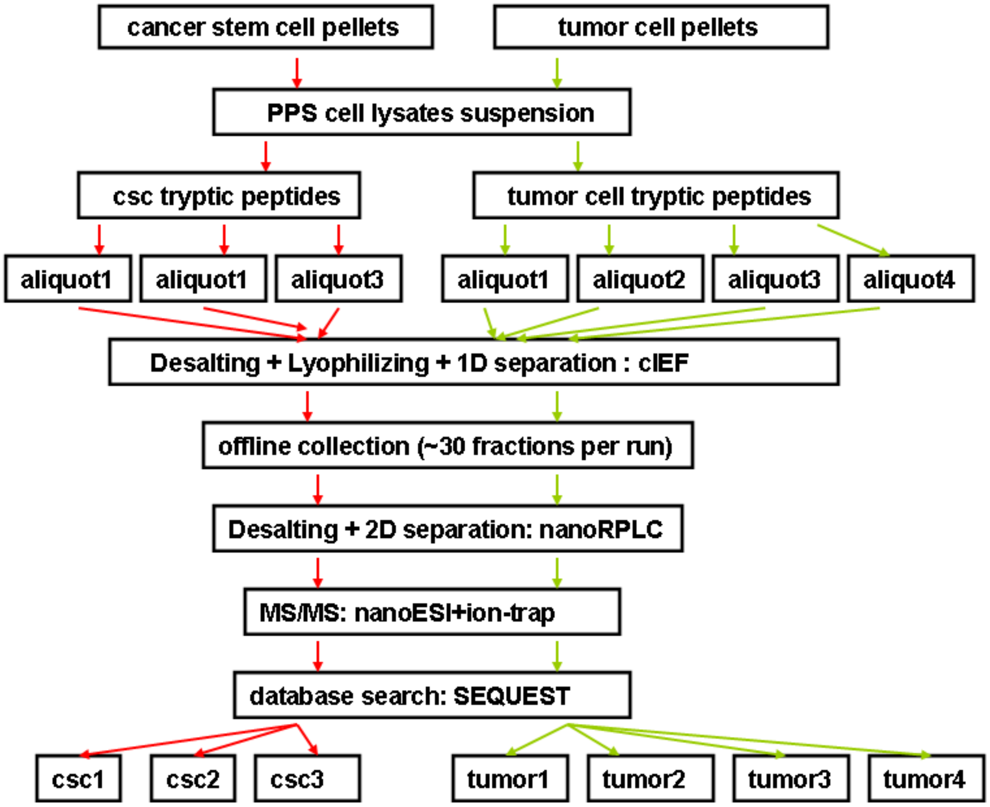
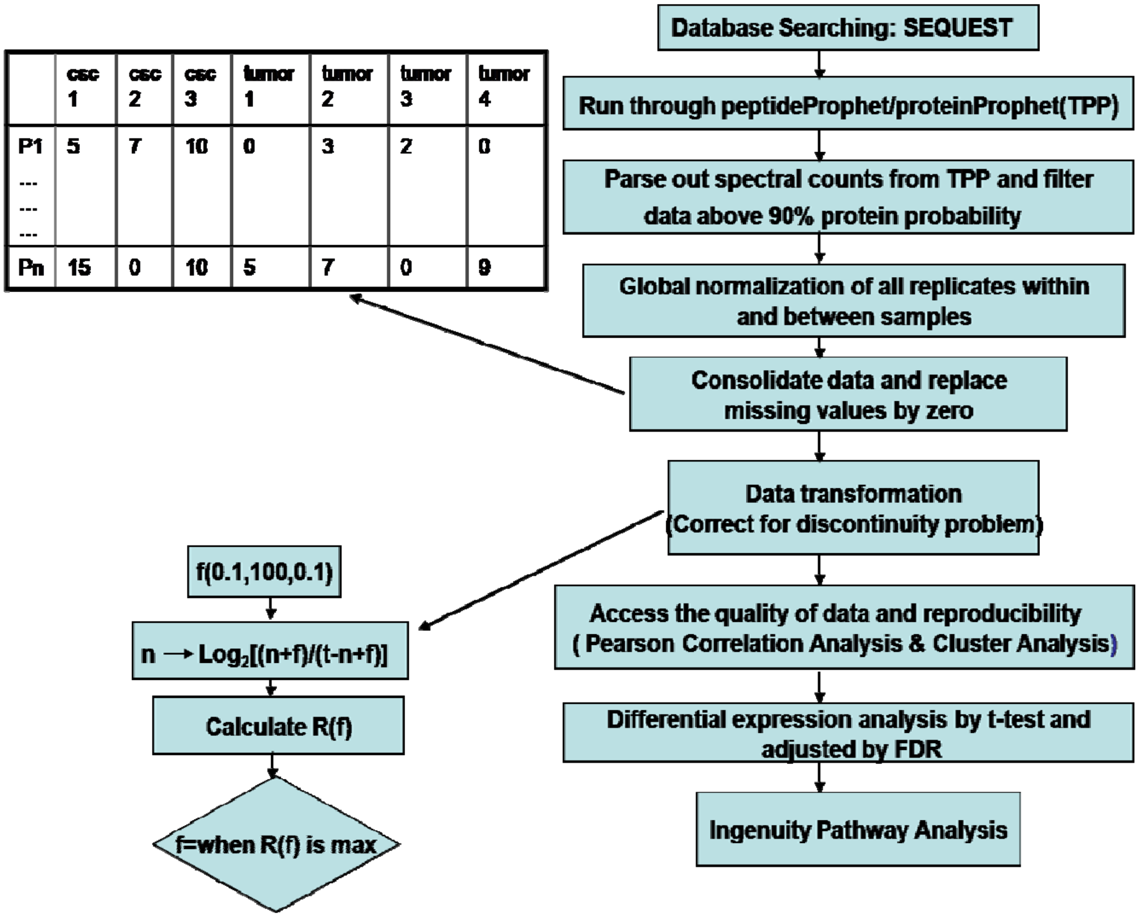


Figure 3.1(b): Data Processing Strategy. Upper left matrix is the consolidated dataset. Lower left flowchart is the correction factor searching scheme and transformation algorithm.



	csc 1	csc 2	csc 3	tumor 1	tumor 2	tumor 3	tumor 4
P1 --- --- ---	5	7	10	0	3	2	0
Pn	15	0	10	5	7	0	9

Database Searching: SEQUEST
Run through peptideProphet/proteinProphet(TPP)
Parse out spectral counts from TPP and filter data above 90% protein probability
Global normalization of all replicates within and between samples
Consolidate data and replace missing values by zero
Data transformation (Correct for discontinuity problem)
Access the quality of data and reproducibility (Pearson Correlation Analysis & Cluster Analysis)
Differential expression analysis by t-test and adjusted by FDR
Ingenuity Pathway Analysis
f(0.1,100,0.1)
n → Log$_2$[(n+f)/(t-n+f)]
Calculate R(f)
f=when R(f) is max

Figure 3.2(a): Theoretical p*I* distribution plot of the first run of CSC group. Fraction number shown in the *X*-axis is plotted against the average of peptides' p*I* value within each fraction shown in the *Y*-axis.

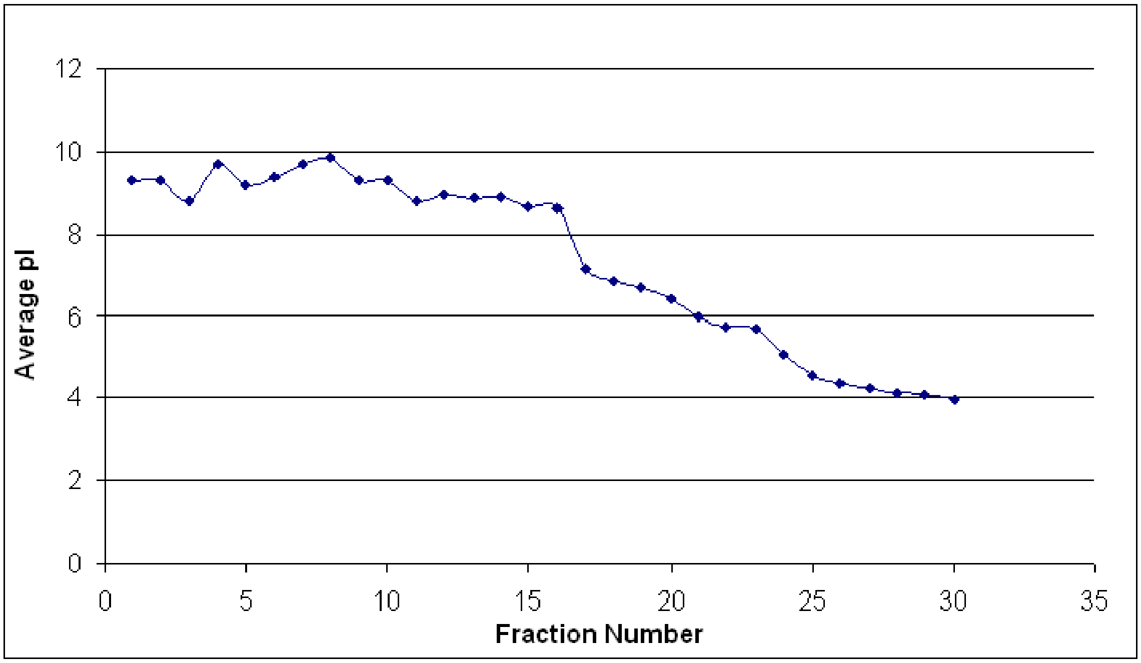


Figure 3.2(b): Distribution of number of identified peptides from each run of tumor group across pI range between 3.5 to 10. X-axis shows their pI value and Y-axis shows the number of identified peptides. Different tumor replicate runs are represented by different colors.

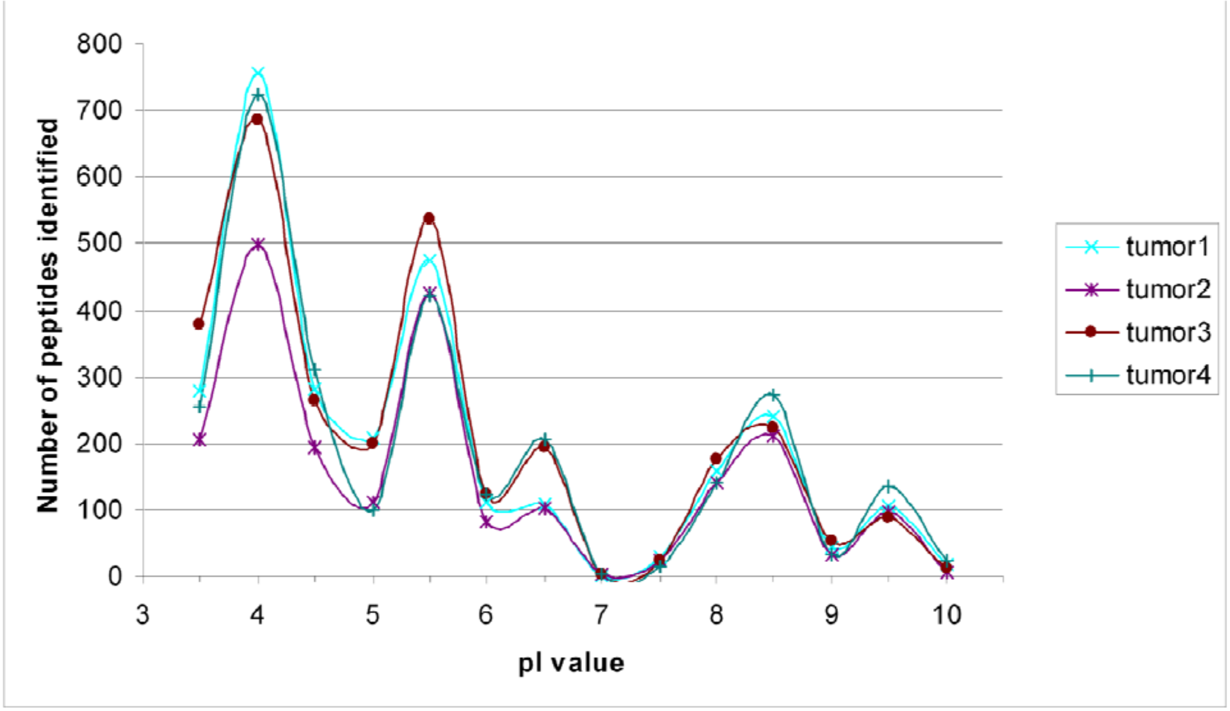


Figure 3.2(c): Pearson correlation plot of all proteins detected with single or more spectral counts in the first and the second run of CSC group.

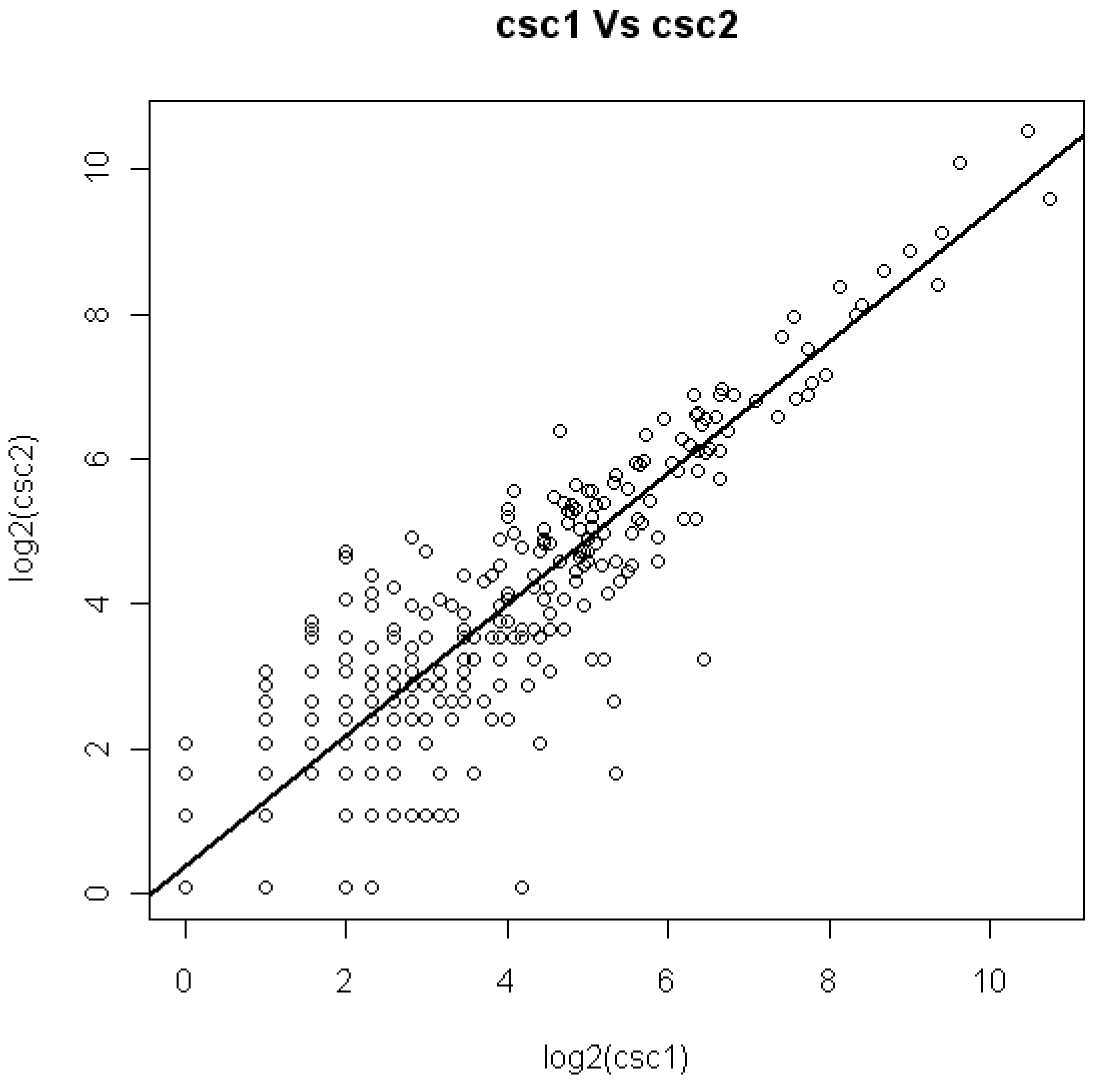


Figure 3.2(d): Pearson correlation plot of all proteins detected with single or more spectral counts in the first and second replicate of tumor group.

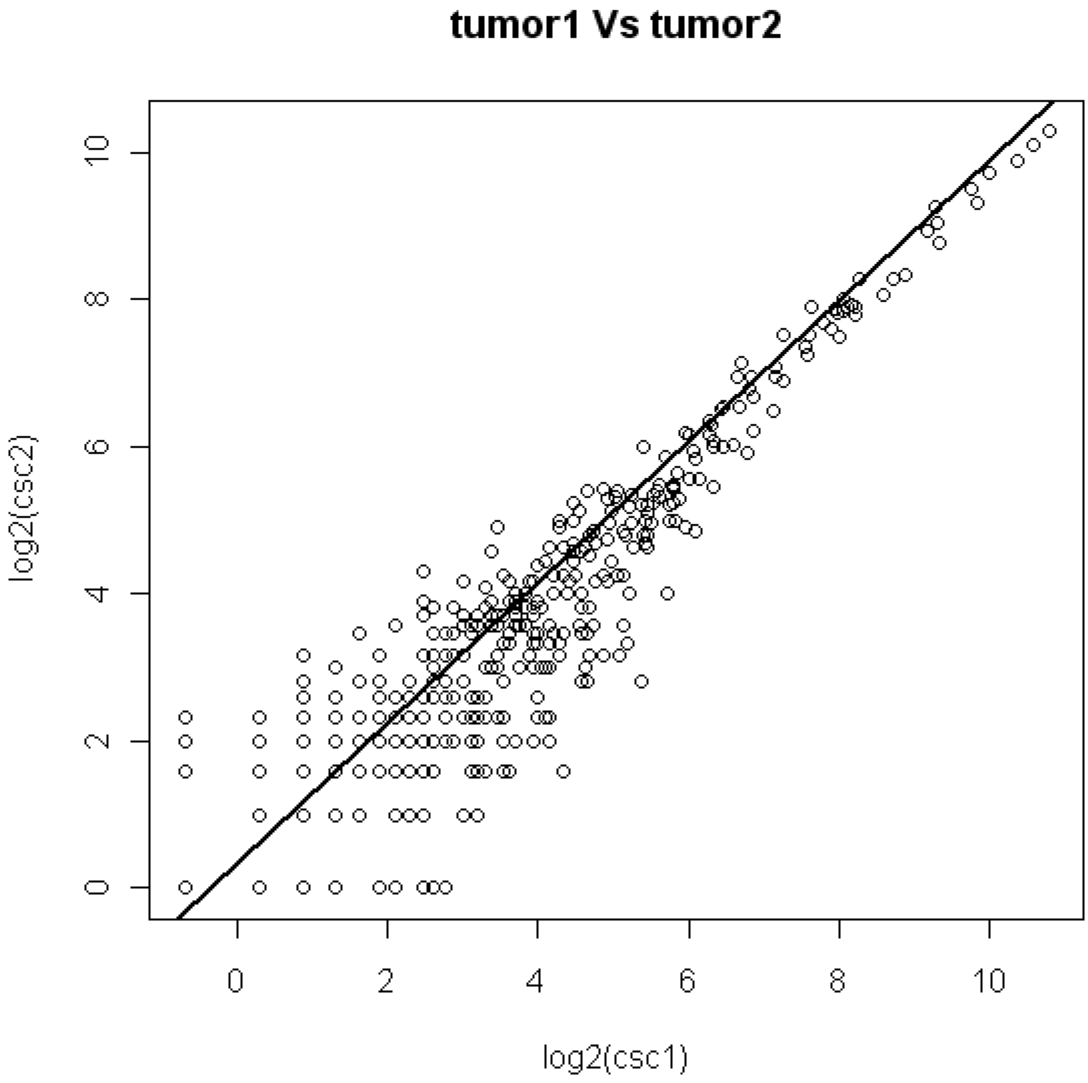


Figure 3.3: Monotonic plot of original data Vs transformed data. Different color and different shapes represent csc1, csc2, csc3, tumor1, tumor2, tumor3, tumor4, respectively. Y-axis represent original data and X-axis represent transformed data on log2 scale.

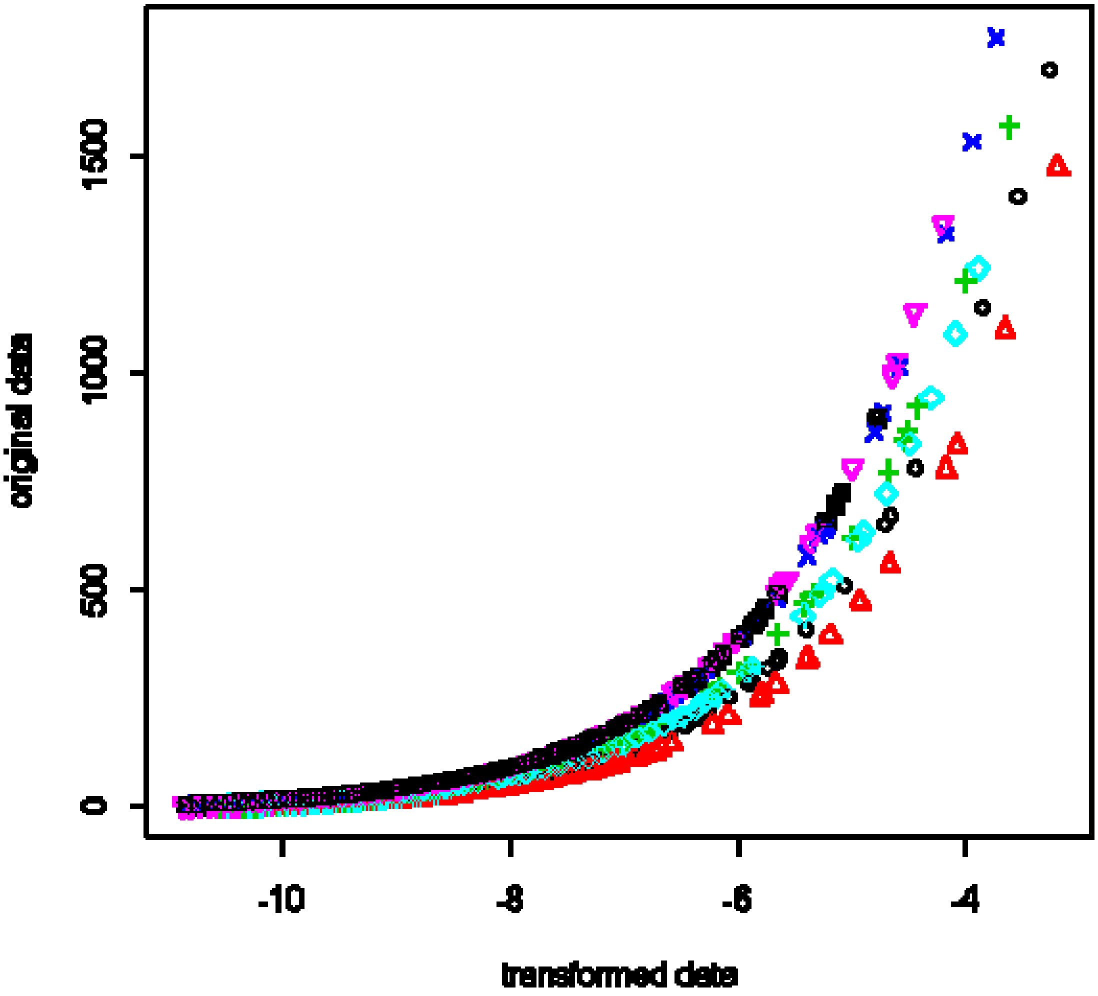


Figure 3.4: Clustering results of the CSC group and tumor group after transforming by f=3.

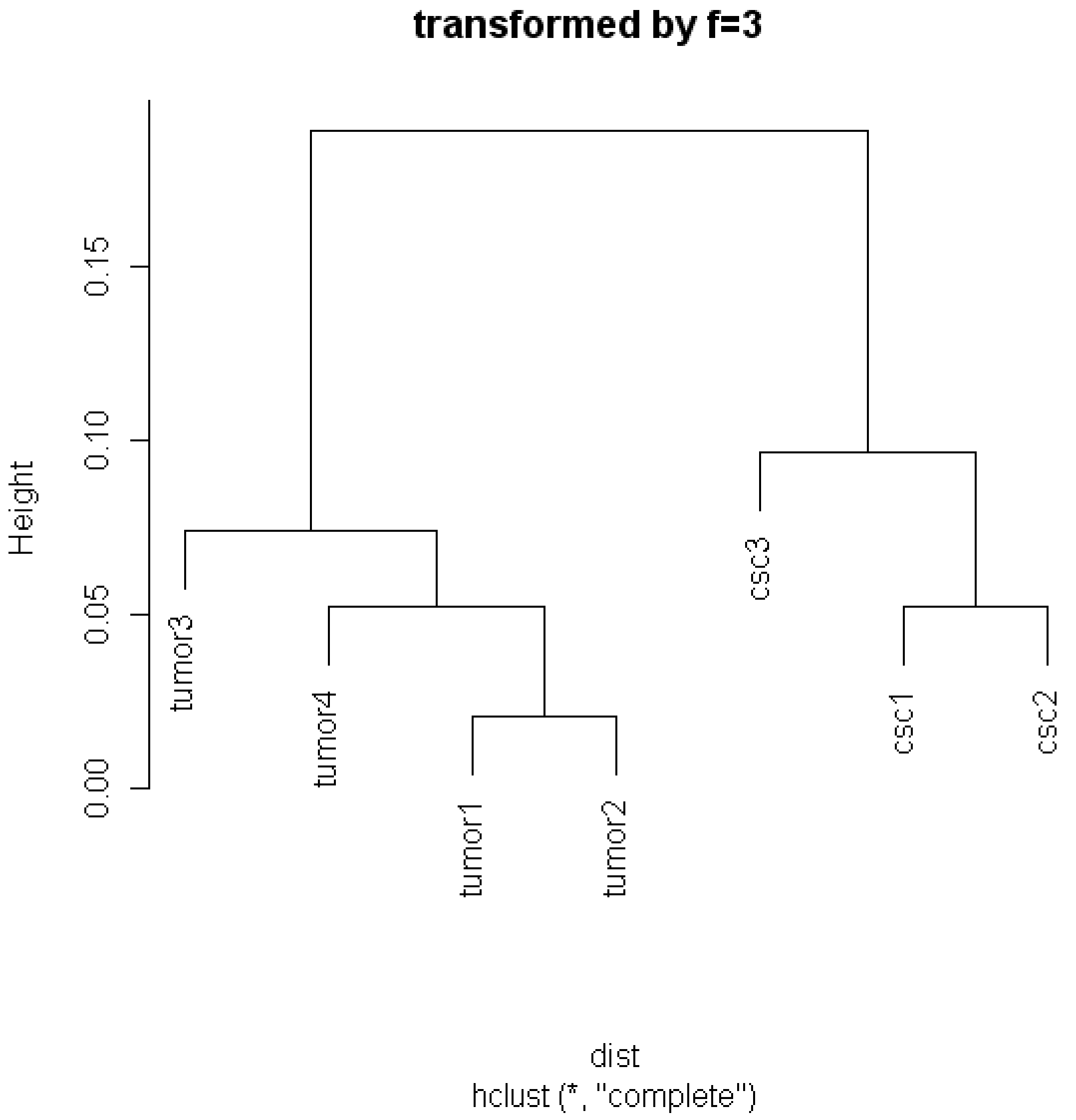


Figure 3.5(a): Cellular Distribution of identified proteins from pooled CSC group.

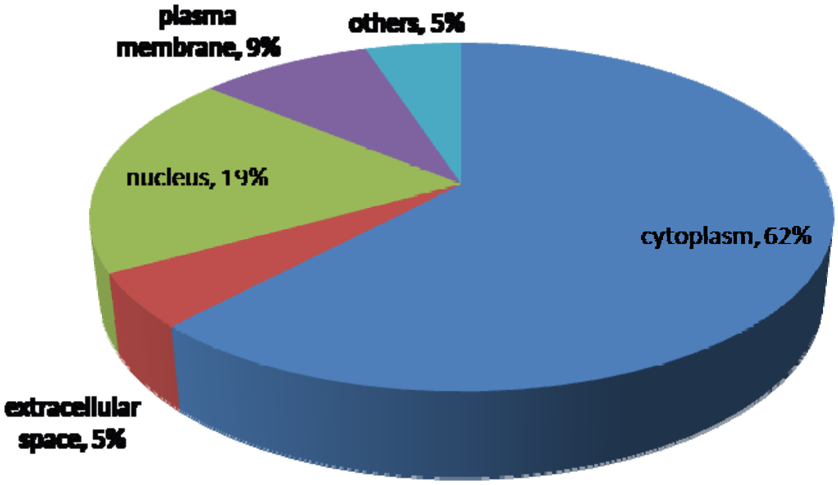


Figure 3.5(b): Cellular Distribution of identified proteins from pooled tumor group.

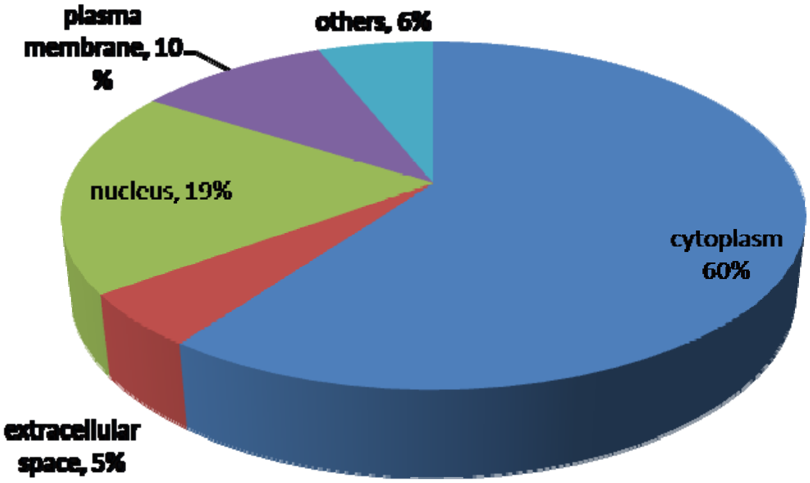


Figure 3.6: Canonical signaling pathways enriched with differentially expressed proteins ranked by significance. A threshold p-value < 0.1 is applied.

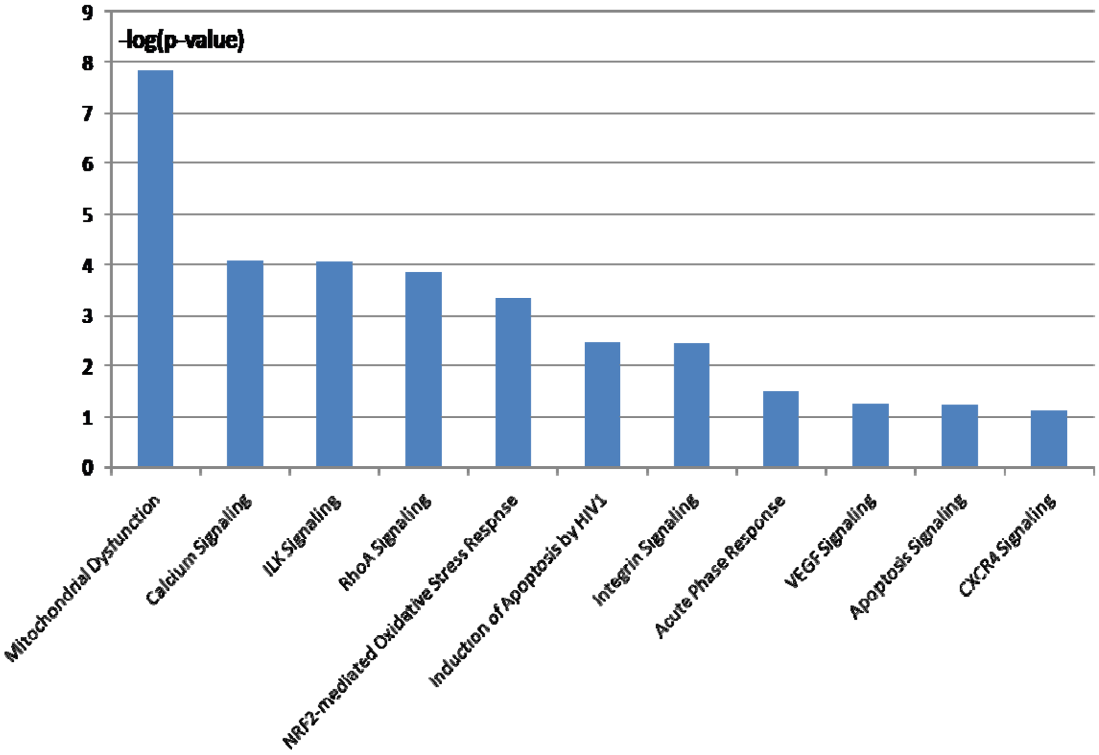
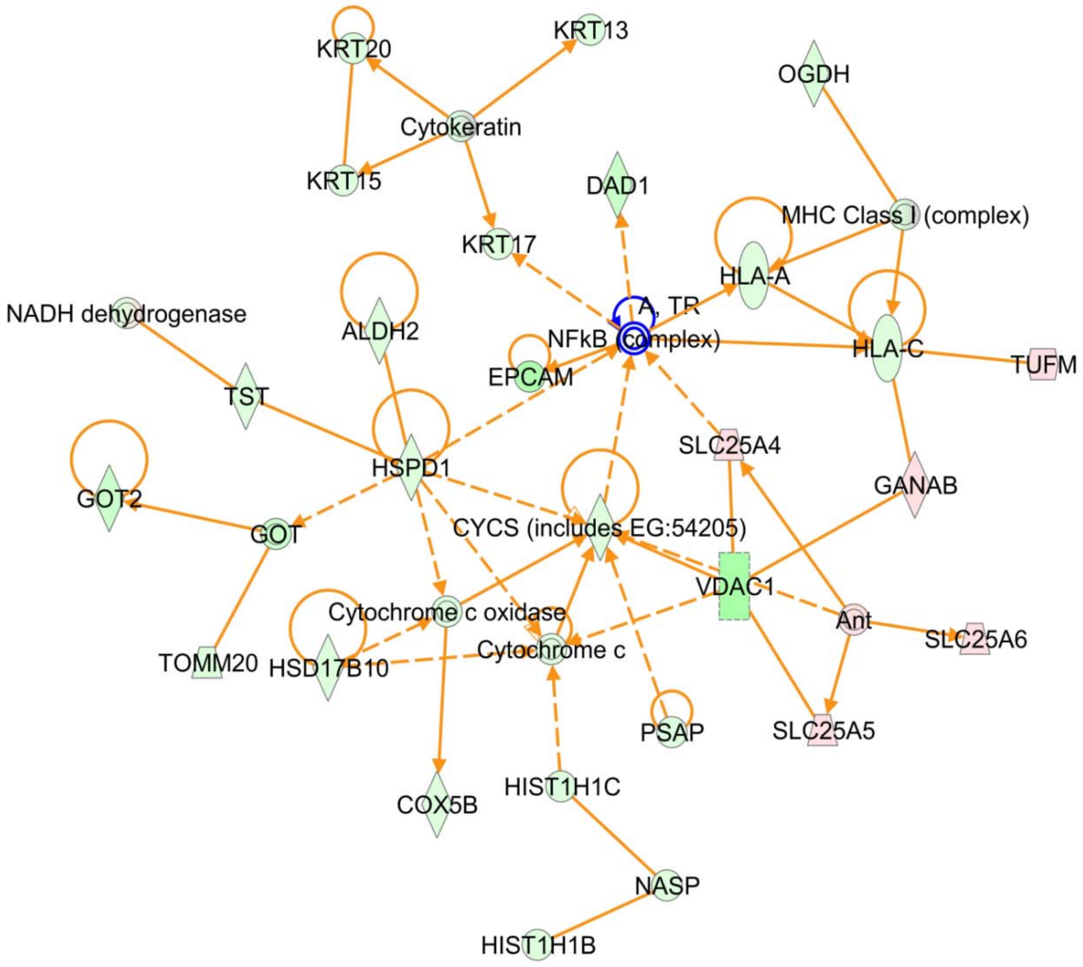


Figure 3.7: The top1 connectivity network constructed by IPA. This network only consists of differentially expressed proteins from experimental data. Red and green circles indicate overexpression and underexpression in the CSC group versus the bulk tumor group, respectively.



CHAPTER 4

DIFFERENTIAL PROFILING STUDIES OF N-LINKED GLYCOPROTEINS IN GLIOBLASTOMA CANCER STEM CELLS UPON TREATMENT WITH GAMMA-SECRETASE INHIBITOR

4.1 Abstract

We have recently demonstrated that Notch pathway blockade by gamma-secretase inhibitor (GSI) depletes cancer stem cells (CSCs) in Glioblastoma Mutiforme (GBM) through reduced proliferation and induced apoptosis. However, the detailed mechanism by which the manipulation of Notch signal induces alterations on post-translational modifications such as glycosylation has not been investigated. Herein, we present a differential profiling work to detect the change of glycosylation pattern upon drug treatment in GBM CSCs. Rapid screening of differential cell surface glycan structures has been performed by lectin microarray on live cells followed by the detection of N-linked glycoproteins from cell lysates using multi-lectin chromatography and label-free quantitative mass spectrometry analysis. A total of 51 and 52 glycoproteins were identified in the CSC and GSI-treated groups, respectively, filtered by a combination of decoy database searching and Trans-Proteomic Pipeline(TPP) processing. Although no significant changes were detected from the lectin microarray experiment, 7 differentially expressed glycoproteins with high confidence were captured after the multi-lectin column including key enzymes involved in glycan processing. Functional annotations of the

altered glycoproteins suggest a phenotype transformation of CSCs toward a less tumorigenic form upon GSI treatment.

4.2 Introduction

The existence of cancer stem cells(CSCs) including in human brain tumors and the implications of promising new therapies that target this small subset of cells have been recently proposed[1-5,8,9]. Notch signaling has been demonstrated as one of the most important molecular mechanisms responsible for CSC properties and we and others have shown that knockdown of this pathway by gamma-secretase inhibitor (GSI) results in attenuated propagation potential of CSCs in GBM[6-9], which is the most aggressive class of brain tumors. However, little is known about the effect on the alteration of glycosylation upon the blockade of Notch signals. Glycoproteins play a critical role in cell-cell recognition events and glycosylation changes have been related to malignant transformation and tumor propagation. Besides the extensively studied role of phosphorylation cascade in Notch signaling, carbohydrate modification has also been shown as an essential regulation mechanism such as the modulation role of O-fucose glycans in Notch receptor function [10-12] and the tight correlation between N-glycosylation and stabilization of Nicastrin which is a component of the gamma-secretase complex[13]. Therefore, it is important to study the changes of glycosylation patterns upon Notch pathway blockade by GSI in GBM in order to better understand the effects of drug treatment.

For the identification of glycoproteins, it is desirable to perform an enrichment step prior to downstream liquid chromatography(LC) coupled mass spectrometry(MS) analysis. The isolation of glycoprotein/glycopeptides is mainly implemented by hydrazide chemistry[14, 15] or affinity capture based on the recognition of different lectins to particular sugar moieties[16] or the combination of both[17]. Lectin-based affinity enrichment on the protein level facilitates potentially multiple glycosylation sites so as to strengthen the relatively weak non-covalent bindings. In addition to the use of a single lectin to capture a particular form of glycan structure, multi-lectin chromatography using lectins with broad specificities has been applied to analyze glycomes in different biological samples[18-21]. In our work, we have utilized three agarose bound lectins: Concanavalin A(ConA), Wheat Germ Agglutinin(WGA) and *Sambucus Ambucus Nigra*(SNA) to produce a broad enrichment of N-linked glycoproteins simultaneously. Another technology for rapid screening of differential glycan structures lies in the development of the lectin microarray, where a large number of lectins are immobilized on a slide to profile the glycoproteins from cell lysates[22] or to obtain cell surface glycan signatures from live cells[23]. Our group has previously demonstrated the feasibility of coupling lectin microarrays for profiling live cells with LC-MS to identify cell surface glycoprotein markers[24].

In the present study, we have employed different strategies to target cell surface glycoproteins and intracellular membrane glycoproteins separately. The profiling of differential cell surface glycan structures has been performed by a fluorescent-assisted lectin microarray with a panel of 16 lectins. A larger scale profiling of N-linked glycoproteins from the soluble fraction of cell lysates has been performed by coupling multi-lectin chromatography with a label-free quantitative MS method. A selective list of differentially expressed glycoproteins has been validated by western blotting assays. The functional relevance of the altered glycosylation patterns has also been inferred and discussed to interpret the biological implications of our findings.

4.3 Materials and Methods

4.3.1 Cell Culture and Treatment

GBM neurosphere cultures were maintained in Neurocult medium (Stem Cell Technologies, Vancouver, BC, Canada, http://www.stemcell.com) supplemented with epidermal growth factor (10 ng/ml) and fibroblast growth factor (10 ng/ml) as previously described[7, 25]. Treatment studies were performed by growing cells in Neurocult medium overnight and replacing the next morning with medium containing γ-secretase inhibitor(Compound E, EMD Chemicals, Gibbstown) dissolved in dimethyl sulfoxide (DMSO) at 1μM.

4.3.2 Lectin Microarray

Sixteen lectins were utilized for the detection of differential cell surface glycan structures as previously described[24]. Briefly, each lectin was printed in three replicates on a SuperAmine slide (Arrayit, Sunnyvale, CA) using a piezoelectric noncontact printer (Nano plotter; GeSiM, GmbH, Germany) and blocked with 1% BSA in PBS (pH7.4) for 1 hr. Fresh GBM CSCs and GSI treated cells were labeled with 10μM CFSE cell-tracing dye (Invitrogen, Carlsbad, CA) and incubated with lectin slides at room temperature for 40min in darkness. After being washed with PBS for 5min, the slides were air-dried and scanned with a microarray scanner (Genepix 4000A; Axon). Genepix 6.0 was used to analyze the images.

4.3.3 Protein Extraction

Cell pellets were resuspended in 1mL of lysis buffer (1% octyl-*β*-D-glucopyranoside, 150 mM NaCl and 1% protease inhibitor mixture (Sigma-Aldrich) in 20 mM Tris-HCl, pH7.4) and homogenized with 60 strokes in a Dounce glass homogenizer with a tight-fitting pestle on ice. The cell lysate was the centrifuged at 40,000*g* for 30 min at 4 °C. Protein concentration from the supernatant was determined by Micro BCA™ Protein Assay Kit (Pierce/Thermo Scientific, Rockford).

4.3.4 Multi-lectin Affinity Chromatography

A single Pierce disposable column was gravity-packed with 1.5mL of agarose-bound ConA, WGA and SNA at 1:1:1 (v/v/v) for individual samples from each biological replicate. The column was first equilibrated with 10 volume of binding buffer (20 mM Tris-HCl, pH7.4, 150 mM NaCl, 1 mM MgCl2, 1mM CaCl2, and 1 mM MnCl2). Cell lysates containing 1mg proteins was diluted four times with binding buffer and passed through the column twice. The column was then washed with 4 volume of binding buffer and eluted with 4 volume of elution buffer (0.2M methyl-a-D-mannopyronoside, 0.2M N-acetyl-glucosamine, 0.2M D-lactose and 0.5M NaCl in 20mM Tris pH 7.4). The Eluent was buffer exchanged with 50mM Ammonia Bicarbonate and concentrated by Microcon YM-10 centrifugal filter devices (10k MWCO) at a final volume of 200µl.

4.3.5 Online nano-RPLC and LTQ Mass Spectrometry

Trypsin digestion was performed by the same protocol described in our previous study[26] prior to online-RPLC (Paradigm MG4 micropump system, Michrom Biosciences Inc., Auburn, CA) connected to a LTQ mass spectrometer (Thermo Finnigan, San Jose, CA). Tryptic digests were reconstituted in 100µl of 5% ACN and 0.3%FA and introduced into a RPLC nano column (3µm x 200 Å, 0.1 mm × 150 mm, C18 AQ particles, Michrom) after a desalting nano trap (300 × 50 mm) (Michrom) with each injection of 20µl. A 2hr linear gradient with 150min from 5 to 40% ACN, 15min from 5% to 80% ACN and another 15min for equilibrium to 5% ACN was used. The remaining LTQ parameters are the same as previously described[27].

4.3.6 Database Searching and Data Processing

MS/MS spectra were searched against Uniprot database using SEUQEST embedded in Proteome Discoverer(version 1.1.0.263). Searching parameters were specified as follows: (1) Fixed modification: carbamidomethylation of Cys residue with a mass shift of 58.1Da; (2) variable modification: oxidation of Met residue with a mass shift of 15.99Da; (3) two missed cleavage sites were allowed; (4) peptide ion mass tolerance: 1.4 Da; (5) fragment ion mass tolerance: 1.0 Da; (6) peptide charges +1, +2, and +3. Searching results were further uploaded to Scaffold(V.2.0) as msf format. Dual filtering criteria for protein identification were employed by combining FDR test from target-decoy database search with a cutoff p-value less than 0.05 and protein/peptide confidence above 95% probability with a minimum of two unique peptides per protein from TPP built in Scaffold. To be noted, employing either target-decoy database searching or TPP alone to filter protein identifications is commonly accepted. Thus, a combination of these two validation strategies represents a more stringent filtering criterion and further increases the identification confidence. Glycoproteins were then confirmed by searching against the post-translational modification annotation in Uniprot database. Spectral counts were parsed out and normalized by a global normalization method.

4.3.7 Western Blot

Western Blot was performed essentially as previously described[26]. Briefly, 20 μg of total proteins from each sample were separated by 4-20% SDS-PAGE and then transferred to PVDF membranes (Bio-Rad, CA). After being blocked for 2 h, the membranes were incubated with antibodies including mouse monoclonal anti-BGAL, mouse polyclonal anti-P4HA1, rabbit polyclonal anti-GANAB, mouse monoclonal anti-beta Actin (Abcam, Cambridge, CA), anti-CATD (BD Transduction Laboratories, Lexington, KY) and anti-THY1 (Abnova, Taibei, China) overnight. After washing three times, the membranes were incubated with HRP conjugated goat anti-rabbit or anti-mouse IgG (H+L) for 1 hr. The blots were visualized with DAB stain (Vector Laboratory, WI).

4.4 Results and Discussion

4.4.1 Detection of Surface Glycoproteins by Lectin Microarray

To probe differential cell surface glycan structures, a panel of 16 lectins covering a wide range of binding specificities to different sugar moieties were printed on the slides and incubated with fluorescent labeled live cells from the CSCs or GSI-treated group. However, no significant intensity changes were detected as a function of the differential expression of cell surface glycans between these two groups. This indicates the total amount of cell surface glycoproteins is not altered after GSI treatment from a macro level point of view, although individual changes may be masked.

4.4.2 Profiling of Intracellular Glycoproteins by Multi-lectin Chromatography

The aforementioned negative results suggest a need of differential glycoprotein profiling on a larger scale. Thus, three widely used lectins (ConA, WGA and SNA) with binding specificities towards α-linked mannose, N-acetylglucosamine and sialic acid were combined to enrich intracellular glycoproteins from the CSCs or GSI-treated group to maximize the coverage. As shown in Figure 4.1, each batch of CSCs or GSI-treated sample were processed in the same way via multi-lectin chromatography and analyzed by LC-MS/MS in triplicates. The whole cell lysates were adjusted to the same amount (1mg) for each biological replicate of each sample prior to lectin enrichment and 10% of the eluent was introduced to the LC-MS/MS for each technical replicate. The dual filtering criteria generated 51 and 52 glycoprotein identifications for the CSC group and the GSI-treated group after database searching. A highly stringent filtering strategy was employed in this study to obtain the most confident identifications. The number of glycoprotein identifications for the CSC group was increased to 88 when the threshold was lowered to a FDR < 0.01 by decoy database searching alone. This result is comparable to a similar study in our group where 73 glycoproteins were identified by using a different combination of lectins (ConA, WGA and PNA) for the group at FDR < 0.01 by decoy database searching[28]. 47 glycoproteins were shared in common and the remainder of the unique identifications may be due to the different glycan-binding specificity of the different lectin (SNA vs PNA) used in each study.

Student's t-test was then used to capture the differentially expressed glycoproteins for each pair with a threshold p-value < 0.05. The differential expression of a protein is accepted when it has a p-value less than 0.05 at least one time out of the three pairs of comparisons and the direction of the changes should be consistent. This results in a total of 27 differentially expressed glycoproteins after GSI treatment as listed in Table 4.1. Seven glycoproteins were considered to be highly confident as they have p-values <0.05

from two pairs of comparisons: beta-galactosidase (BGAL), Calumenin(CALU), Deoxyribonuclease-2-alpha(DNS2A), Neural alpha-glucosidase AB(GANAB), Hypoxia up-regulated protein1(HYOU1), Prolyl 4-hydroxylase subunit alpha-1(P4HA1) and Serotransferrin(TRFE).

To compare the enrichment efficiency, a fraction of the whole cell lysates from the third batch was processed in the same manner excluding the step of multi-lectin chromatography and it was analyzed by LC-MS/MS in triplicates with each injection of ~1µg. Figure 4.2 illustrates a comparison of the spectral counts assigned to each highly confident glycoprotein before and after multi-lectin chromatography (Note: DNS2A is not detected from batch 3). The averaged spectral counts for each glycoprotein are significantly increased up to 17 fold after the enrichment. Protein Disulfide-isomerase (PDIA) shown in the bottom of this graph is not a glycoprotein, however, binding of P4HA1 and PDIA has been previously reported[29]. Thus, the increase of PDIA after enrichment is expected as a consequence of protein-protein interaction.

4.4.3 Verification of Differential Expression by Western Blot

Five of the differentially expressed glycoproteins were selected for verification by Western Blot: GANAB, BGAL, P4HA1, CATD and THY1. The fold changes and biological functions are shown in Table 4.2. The expression levels of the first four proteins are down-regulated after GSI treatment, whereas THY1 exhibits increased expression. The intensities of the bands are not quantitative results to compare the expression levels across different proteins. However, the lowest intensity detected for THY1 correlates well with the label-free MS results where this protein is assigned with the least spectral counts compared to others.

4.4.4 Protein-Protein Interaction Network

The functional relevance of the differentially expressed glycoproteins was searched against the STRING database which enables public access to retrieve protein-protein interactions[30]. Nine glycoproteins are functionally linked to each other with medium to high confidence as displayed in Figure 4.3. They are grouped according to their sub-cellular localizations and the thickness of an edge positively correlates with the level of association. The links between CTSD, GLB1, CTSA, HEXA and HEXB are of high confidence possibly due to their co-localization in the lysosome compartment. The strongest association is assigned between GLB1 and CTSA which have been demonstrated to interact to form the lysosomal multienzyme complex[31].

4.5 Discussion

4.5.1 Multi-lectin Affinity Strategy

In addition to N-linked glycosylated proteins, non-glycoproteins were also detected by mass spectrometry analysis after glycoprotein enrichment. The identifications of non-glycosylated proteins are mainly due to non-specific bindings to the multi-lectin column resulting from protein-protein interactions. 1% octyl-β-D-glucopyranoside was used in our lysis procedure where this non-ionic detergent has been demonstrated to be effective in releasing a wide range of membrane proteins and is compatible with downstream MS analysis[32]. The detergent is considered to be a mild surfactant which may not disrupt strong direct bindings between proteins. Moreover, the detergent concentration in the sample is diluted before loading onto the multi-lectin column to enable the non-covalent binding and the elution buffer used subsequently is under physiological condition with pH 7.4 which further preserves the formation of the endogenous protein complex. Thus, the detection of non-glycosylated proteins can be explained as a result of co-precipitation in the enrichment step. For example, Ribophorin 1 is a glycoprotein detected with reduced expression level after treatment. A non-glycosylated protein Ubiqilin 4 is also detected with reduced level possibly due to its demonstrated binding to Ribophorin 1[33]. The elution of a larger protein complex resulting from co-precipitation occurs in the case where the glycoprotein P4HA1 is the primary target. P4HA1 binds to PDIA exhibiting direct interaction with 1433G[34] which further binds to three other proteins. Future improvements on optimization of the lysis buffer or washing step condition by carefully increasing the detergent strength will be beneficial to reduce the number of off-target identifications. Also, the identification of glycoproteins is based on the information from the Uniprot database as the mass accuracy of the LTQ mass spectrometer is not sufficient to distinguish a 1Da mass unit shift from PNGaseF cleavage, which is commonly used to identify the glycosylation site by high resolution mass spectrometers. The incorporation of isotope labeling by inducing a larger mass shift that can be recognized by the LTQ mass spectrometer will be helpful to discover novel glycosylation sites in future studies.

4.5.2 Important Differentially Expressed Glycoproteins

Five differentially expressed glycoproteins have been validated by Western blot. GANAB is the α-subunit of an important ER resident enzyme Glucosidase II which sequentially cleaves inner α-1,3-linked glucose residues from N-linked oligosaccharides on nascent glycoproteins[35] and it has been previously reported to be purified by ConA[36]. The decreased expression after GSI treatment may suggest altered glycan processing effects on the maturation of glycoproteins. BGAL is an essential enzyme that catalyzes the hydrolysis of β-galactosides into monosaccharides. A previous report shows it can be purified from either a ConA or WGA lectin column[37]. The potential role in GBM CSCs remains unknown.

P4HA1 is another key enzyme in collagen synthesis which catalyzes the formation of 4-hydroxyproline and it exhibits binding to ConA[38]. Decreased P4HA1 expression has been observed after knocking down its transcriptional regulator Hypoxia-induced-factor 1-alpha(H1F-1α) in GBM CSCs[39]. In addition, the translocation of H1F-1α to bind Notch ICD and the convergence required to maintain stem cell properties have been proposed[40]. Therefore, the similar down-regulated pattern of P4HA1 after GSI treatment may be due to the blockade of Notch and the reduced level of Notch ICD.

CATD which is an essential lysosomal aspartyl endopeptidase is also detected to be down-regulated after treatment. It has been reported that the N-linked oligosaccharide chains consists of a cluster of mannose 6-sulfate residues[41] enabling the purification of this protein by ConA[42]. Over-expression of CATD stimulates breast cancer tumorigenicity and metastasis[43]. It has been reported to play an essential role at multiple breast cancer progression steps by promoting cell proliferation and angiogenesis via inhibition of tumor apoptosis[44]. Therefore, the reduced level of CATD after treatment may suggest a transformation of the phenotype towards a less tumorigenic form. This hypothesis is also in line with the detection of increased THY1(also known as CD90) after treatment which points to the same inference. THY1 is a heavily N-glycosylated cell surface antigen with ConA binding sites[45]. Previous studies proposed that THY1 is a putative tumor suppressor gene in ovarian cancer[46, 47] and nasopharyngeal carcinoma[48]. The protein expression level was found to be exclusively in the non-tumorigenic models as compared to their tumorigenic counterparts [46-48]. However, the correlation between the expression level of THY1 and the degree of tumorigenicity remains controversial. For example, higher expression of THY1 at the mRNA level is found in CD133$^+$ GBM cells compared to CD133$^-$ cells[49]; THY1 positive cells sorted from hepatocellular carcinoma cell lines exhibit tumorigenecity rather than their counterparts. In our previous studies, the protein expression level of THY1 is found to be higher in GBM CSCs than normal cells[24] and no difference is observed when compared to the differentiated GBM cells[28]. Therefore, the correlation between THY1 and tumorigenicity may vary depending on the particular type of cancer, the degree of tumor grade and the specific perturbation treatment. Moreover, it is unknown that if the correlation relies on the total protein level of THY1, the glycosylation level of THY1, or even a specific glycoform level of THY1. Future investigations focusing on such detailed characterizations will help us clarify the uncertainty of the impact on tumorigenicity mediated by THY1.

In addition to the above glycoproteins which have been validated by western blot, another down-regulated glycoprotein PGCB(Brevican) detected in our study after treatment appears to be particularly interesting. PGCB is a brain specific proteoglycan with one of the isoforms oversialylated[50] and the interaction with SNA may strengthen its binding to the multi-lectin column. The critical role of PGCB in promoting GBM dispersion has been demonstrated in a manner that over-expression of this glycoprotein increases tumor invasion whereas knockdown inhibits it[51, 52], which further supports our hypothesis that a phenotype transformation towards a less tumorigenic form occurs upon drug treatment.

4.5.3 Exploration of GLMM

GLMM is an extension to the generalized linear model where random effects are incorporated in addition to fixed effects in the linear predictors. The strength of such mixed-effects models lies in their capability of providing a powerful and flexible analysis of correlated observations from a nested experimental design. In the present study, the dependency results from the splitting the same source into two parts: with (the CSC group) and without treatment (the GSI group). Thus, classical statistical testing which assumes independency may not be suited. Another advantage is that two levels of variations (between- and within-group) are taken care of simultaneously by the predictors. Therefore, the use of such mixed-effects models makes it possible to analyze the variability of hierarchical data structure from biological experiments. The application of GLMM to analyze the differential expression of glycoprotein data is explored here.

The matrix form of GLMM is expressed as follows:

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i \qquad \text{Eq.1}$$

where Y denotes the spectral counts and is assumed to follow Poisson distribution; X denotes the fixed effect/group effect indicating the between-group variations; Z denotes the random effects indicating the within-group variations; ε denotes the error residues. A larger dataset than previously described in section 4.3 is tested. Specifically, each group has 3 biological replicates and each biological replicate has 3, 5 and 7 technical replicates, respectively. Thus, a total of 15 observations are collected for each protein. The spectral count for each protein is modeled as a linear regression of the fixed effect and the random effect. The model is fitted by maximizing the REML function. The fixed effect and random effect are then assessed by testing the significance of their corresponding coefficients. This process is implemented by the R-programming package (lme4 library) with the built-in function (glmer). Proteins with p-values less than 0.05 for the coefficients of fixed effect (β) are considered to have significantly differential expression levels. The output shows one of the glycoprotein GANAB which has been confirmed by the western blot experiments as aforementioned is differentially expressed with the p-value of 0.002.

4.6 Conclusion

In this work, we investigate the alteration of glycosylation pattern upon treatment of GSI in GBM CSCs. We have utilized a combination of lectin microarray and muti-lectin chromatography coupled RPLC-MS analysis to target cell surface glycoproteins and glycoproteins from cell lysates, respectively. While no significant changes have been detected from microarray screening, several differentially expressed intracellular membrane proteins and plasma membrane proteins were captured by the multi-lectin enrichment approach. The finding of down-regulation of GANAB and BGAL may suggest an altered glycan processing while reduced level of CATD and increased expression of THY1 may imply attenuated proliferation and elevated apoptosis upon GSI treatment. Future improvements may involve the optimization of detergent condition and incorporation of isotope labeling to increase the percentage of glycoprotein identifications, as well as the optimization of GLMM model. Overall, our study provides information regarding the influence of GSI treatment on glycosylation in GBM CSCs which may lead to an improved understanding of drug mechanism.

4.7 References

1. Vescovi AL, G.R., Reynolds BA., *Brain tumour stem cells.* Nat Rev Cancer, 2006. **6**(6): p. 425-36.
2. Read TA, F.M., Markant SL, McLendon RE, Wei Z, Ellison DW, Febbo PG, Wechsler-Reya RJ., *Identification of CD15 as a marker for tumor-propagating cells in a mouse model of medulloblastoma.* Cancer Cell, 2009. **15**(2): p. 135-47.
3. Ward RJ, L.L., Graham K, Satkunendran T, Yoshikawa K, Ling E, Harper L, Austin R, Nieuwenhuis E, Clarke ID, Hui CC, Dirks PB., *Multipotent CD15+ cancer stem cells in patched-1-deficient mouse medulloblastoma.* Cancer Res, 2009. **69**(11): p. 4682-90.
4. Son MJ, W.K., Nam DH, Lee J, Fine HA., *SSEA-1 is an enrichment marker for tumor-initiating cells in human glioblastoma.* Cell Stem Cell, 2009. **4**(5): p. 440-52.
5. Gilbert CA, R.A., *Cancer stem cells: cell culture, markers, and targets for new therapies.* J Cell Biochem, 2009. **108**(5): p. 1031-8.
6. Stockhausen MT, K.K., Poulsen HS., *The functional role of Notch signaling in human gliomas.* Neuro Oncol, 2010. **12**(2): p. 199-211.
7. Fan X, K.L., Zhu TS, Soules ME, Talsma CE, Gul N, Koh C, Zhang J, Li YM, Maciaczyk J, Nikkhah G, Dimeco F, Piccirillo S, Vescovi AL, Eberhart CG., *NOTCH pathway blockade depletes CD133-positive glioblastoma cells and inhibits growth of tumor neurospheres and xenografts.* Stem Cells, 2010. **28**(1): p. 5-16.
8. Fan X, M.W., Khaki L, Stearns D, Chun J, Li YM, Eberhart CG., *Notch pathway inhibition depletes stem-like cells and blocks engraftment in embryonal brain tumors.* Cancer Res, 2006. **66**(15): p. 7445-52.
9. Fan X, E.C., *Medulloblastoma stem cells.* J Clin Oncol, 2008. **26**(17): p. 2821-7.
10. P., S., *Regulation of Notch signaling by glycosylation.* Curr Opin Struct Biol, 2007. **17**(5): p. 530-5.
11. Jafar-Nejad H, L.J., Fernandez-Valdivia R., *Role of glycans and glycosyltransferases in the regulation of Notch signaling.* Glycobiology, 2010. **20**(8): p. 931-49.
12. Takeuchi H, H.R., *Role of glycosylation of Notch in development.* Semin Cell Dev Biol., 2010. **21**(6): p. 638-45.
13. Tomita T, K.R., Takikawa R, Iwatsubo T., *Complex N-glycosylated form of nicastrin is stabilized and selectively bound to presenilin fragments.* FEBS Lett, 2002. **520**(1-3): p. 117-21.
14. Zhang H, L.X., Martin DB, Aebersold R., *Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry.* Nat Biotechnol, 2003. **21**(6): p. 660-6.

15. Chen R, J.X., Sun D, Han G, Wang F, Ye M, Wang L, Zou H., *Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry.* J Proteome Res, 2009. **8**(2): p. 651-61.

16. Xiong L, A.D., Regnier F., *Comparative proteomics of glycoproteins based on lectin selection and isotope coding.* J Proteome Res, 2003. **8**(2): p. 651-61.

17. Kaji H, S.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): p. 667-72.

18. Yang Z, H.W., *Approach to the comprehensive analysis of glycoproteins isolated from human serum using a multi-lectin affinity column.* J Chromatogr A, 2004. **1053**(1-2): p. 79-88.

19. Yang Z, H.W., *Monitoring glycosylation pattern changes of glycoproteins using multi-lectin affinity chromatography.* J Chromatogr A, 2005. **1070**(1-2): p. 57-64.

20. Orazine CI, H.M., Hancock WS, Hattersley M, Hanke JH., *A proteomic analysis of the plasma glycoproteins of a MCF-7 mouse xenograft: a model system for the detection of tumor markers.* J Proteome Res, 2008. **7**(4): p. 1542-54.

21. Lee HJ, N.K., Choi EY, Kim KS, Kim H, Paik YK., *Simple method for quantitative analysis of N-linked glycoproteins in hepatocellular carcinoma specimens.* J Proteome Res, 2010. **9**(1): p. 308-18.

22. Pilobello KT, S.D., Mahal LK., *Proc Natl Acad Sci U S A.* A ratiometric lectin microarray approach to analysis of the dynamic mammalian glycome, 2007. **104**(28): p. 11534-9.

23. Tao SC, L.Y., Zhou J, Qian J, Schnaar RL, Zhang Y, Goldstein IJ, Zhu H, Schneck JP., *Tao SC, Li Y, Zhou J, Qian J, Schnaar RL, Zhang Y, Goldstein IJ, Zhu H, Schneck JP.* Glycobiology, 2008. **18**(10): p. 761-9.

24. He J, L.Y., Xie X, Zhu T, Soules M, DiMeco F, Vescovi AL, Fan X, Lubman DM., *Identification of cell surface glycoprotein markers for glioblastoma-derived stem-like cells using a lectin microarray and LC-MS/MS approach.* J Proteome Res, 2010. **9**(5): p. 2565-72.

25. Galli R, B.E., Orfanelli U, Cipelletti B, Gritti A, De Vitis S, Fiocco R, Foroni C, Dimeco F, Vescovi A., *Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma.* Cancer Res, 2004. **64**(19): p. 7011-21.

26. Dai L, L.C., Shedden KA, Misek DE, Lubman DM., *Comparative proteomic study of two closely related ovarian endometrioid adenocarcinoma cell lines using cIEF fractionation and pathway analysis.* Electrophoresis, 2005. **30**(7): p. 1119-31.

27. Dai L, L.C., Shedden KA, Lee CJ, Li C, Quoc H, Simeone DM, Lubman DM., *Quantitative proteomic profiling studies of pancreatic cancer stem cells.* J Proteome Res, 2010. **9**(7): p. 3394-402.

28. He J, L.Y., Zhu TS, Xie X, Costello MA, Talsma CE, Flack CG, Crowley JG, Dimeco F, Vescovi AL, Fan X, Lubman DM., *Glycoproteomic Analysis of Glioblastoma Stem Cell Differentiation.* J Proteome Res, 2010.

29. Kukkola L, H.R., Kivirikko KI, Myllyharju J., *Identification and characterization of a third human, rat, and mouse collagen prolyl 4-hydroxylase isoenzyme.* J Biol Chem, 2003. **278**(48): p. 47685-93.

30. Jensen LJ, K.M., Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C., *STRING 8--a global view on proteins and their functional interactions in 630 organisms.* Nucleic Acids Res, 2009. **37**(Database issue): p. D412-6.

31. Santamaria R, C.A., Callahan JW, Grinberg D, Vilageliu L., *Expression and characterization of 14 GLB1 mutant alleles found in GM1-gangliosidosis and Morquio B patients.* J Lipid Res, 2007. **48**(10): p. 2275-82.

32. McDonald CA, Y.J., Marathe V, Yen TY, Macher BA., *Combining results from lectin affinity chromatography and glycocapture approaches substantially improves the coverage of the glycoproteome.* Mol Cell Proteomics, 2009. **8**(2): p. 287-301.

33. Lim J, H.T., Shaw C, Patel AJ, Szabó G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabási AL, Vidal M, Zoghbi HY., *protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration.* Cell, 2006. **125**(4): p. 801-14.

34. Jin J, S.F., Stark C, Wells CD, Fawcett JP, Kulkarni S, Metalnikov P, O'Donnell P, Taylor P, Taylor L, Zougman A, Woodgett JR, Langeberg LK, Scott JD, Pawson T., *Proteomic, functional, and domain-based analysis of in vivo 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization.* Curr Biol, 2004. **14**(16): p. 1436-50.

35. Pelletier MF, M.A., Sevigny G, Jakob CA, Tessier DC, Chevet E, Menard R, Bergeron JJ, Thomas DY., *The heterodimeric structure of glucosidase II is required for its activity, solubility, and localization in vivo.* Glycobiology, 2000. **10**(8): p. 815-27.

36. Martiniuk F, E.A., Hirschhorn R., *Identity of neutral alpha-glucosidase AB and the glycoprotein processing enzyme glucosidase II. Biochemical and genetic studies.* J Biol Chem, 1985. **260**(2): p. 1238-42.

37. Heyworth CM, W.C., *The binding of human liver acid beta-galactosidase to wheat-germ lectin is influenced by aggregation state of the enzyme.* Biochem J, 1982. **201**(3): p. 615-9.

38. Alvarez-Manilla G, W.N., Atwood J 3rd, Orlando R, Dalton S, Pierce M., *Glycoproteomic analysis of embryonic stem cells: identification of potential glycobiomarkers using lectin affinity chromatography of glycopeptides.* J Proteome Res, 2010. **9**(5): p. 2062-75.

39. Méndez O, Z.J., Esencay M, Lukyanov Y, Santovasi D, Wang SC, Newcomb EW, Zagzag D., *Knock down of HIF-1alpha in glioma cells reduces migration in vitro and invasion in vivo and impairs their ability to form tumor spheres.* Mol Cancer, 2010. **9**: p. 133.

40. Gustafsson MV, Z.X., Pereira T, Gradin K, Jin S, Lundkvist J, Ruas JL, Poellinger L, Lendahl U, Bondesson M., *Hypoxia requires notch signaling to maintain the undifferentiated cell state.* Dev Cell, 2005. **9**(5): p. 617-28.

41. Journet A, C.A., Jehan S, Adessi C, Freeze H, Klein G, Garin J., *Characterization of Dictyostelium discoideum cathepsin D.* J Cell Sci, 1999. **112**(pt21): p. 3833-43.

42. Srivastava PN, N.V., *Isolation of rabbit testicular cathepsin D and its role in the activation of proacrosin.* Biochem Biophys Res Commun, 1982. **109**(1): p. 63-9.

43. Liaudet-Coopman E, B.M., Derocq D, Garcia M, Glondu-Lassis M, Laurent-Matha V, Prébois C, Rochefort H, Vignon F., *Cathepsin D: newly discovered functions of a long-standing aspartic protease in cancer and apoptosis.* Cancer Lett, 2006. **237**(2): p. 167-79.

44. Berchem G, G.M., Gleizes M, Brouillet JP, Vignon F, Garcia M, Liaudet-Coopman E., *Cathepsin-D affects multiple tumor progression steps in vivo: proliferation, angiogenesis and apoptosis.* Oncogene, 2002. **21**(38): p. 5951-5.

45. K., T., *Heterogeneity of epidermal Thy-1-positive cells defined by lectin-binding sites.* J Invest Dermatol, 1986. **86**(3): p. 222-5.

46. Abeysinghe HR, C.Q., Xu J, Pollock S, Veyberman Y, Guckert NL, Keng P, Wang N., *THY1 expression is associated with tumor suppression of human ovarian cancer.* Cancer Genet Cytogenet, 2003. **143**(2): p. 125-32.

47. Abeysinghe HR, P.S., Guckert NL, Veyberman Y, Keng P, Halterman M, Federoff HJ, Rosenblatt JP, Wang N., *The role of the THY1 gene in human ovarian cancer suppression based on transfection studies.* Cancer Genet Cytogenet, 2004. **149**(1): p. 1-10.

48. Lung HL, B.D., Xie D, Cheung AK, Cheng Y, Kumaran MK, Miller L, Liu ET, Guan XY, Sham JS, Fang Y, Li L, Wang N, Protopopov AI, Zabarovsky ER, Tsao SW, Stanbridge EJ, Lung ML., *THY1 is a candidate tumour suppressor gene with decreased expression in metastatic nasopharyngeal carcinoma.* Oncogene, 2005. **24**(43): p. 6525-32.

49. Liu G, Y.X., Zeng Z, Tunici P, Ng H, Abdulkadir IR, Lu L, Irvin D, Black KL, Yu JS., *Analysis of gene expression and chemoresistance of CD133+ cancer stem cells in glioblastoma.* Mol Cancer, 2005. **2**(5): p. 67.

50. Viapiano MS, B.W., Piepmeier J, Hockfield S, Matthews RT., *Novel tumor-specific isoforms of BEHAB/brevican identified in human malignant gliomas.* Cancer Res, 2005. **65**(15): p. 6726-33.

51. Viapiano MS, H.S., Matthews RT., *BEHAB/brevican requires ADAMTS-mediated proteolytic cleavage to promote glioma invasion.* J Neurooncol, 2008. **88**(3): p. 261-72.

52. Hu B, K.L., Matthews RT, Viapiano MS., *The proteoglycan brevican binds to fibronectin after proteolytic cleavage and promotes glioma cell motility.* J Biol Chem, 2008. **283**(36): p. 24848-59.

Figure 4.1: Experimental work-flow

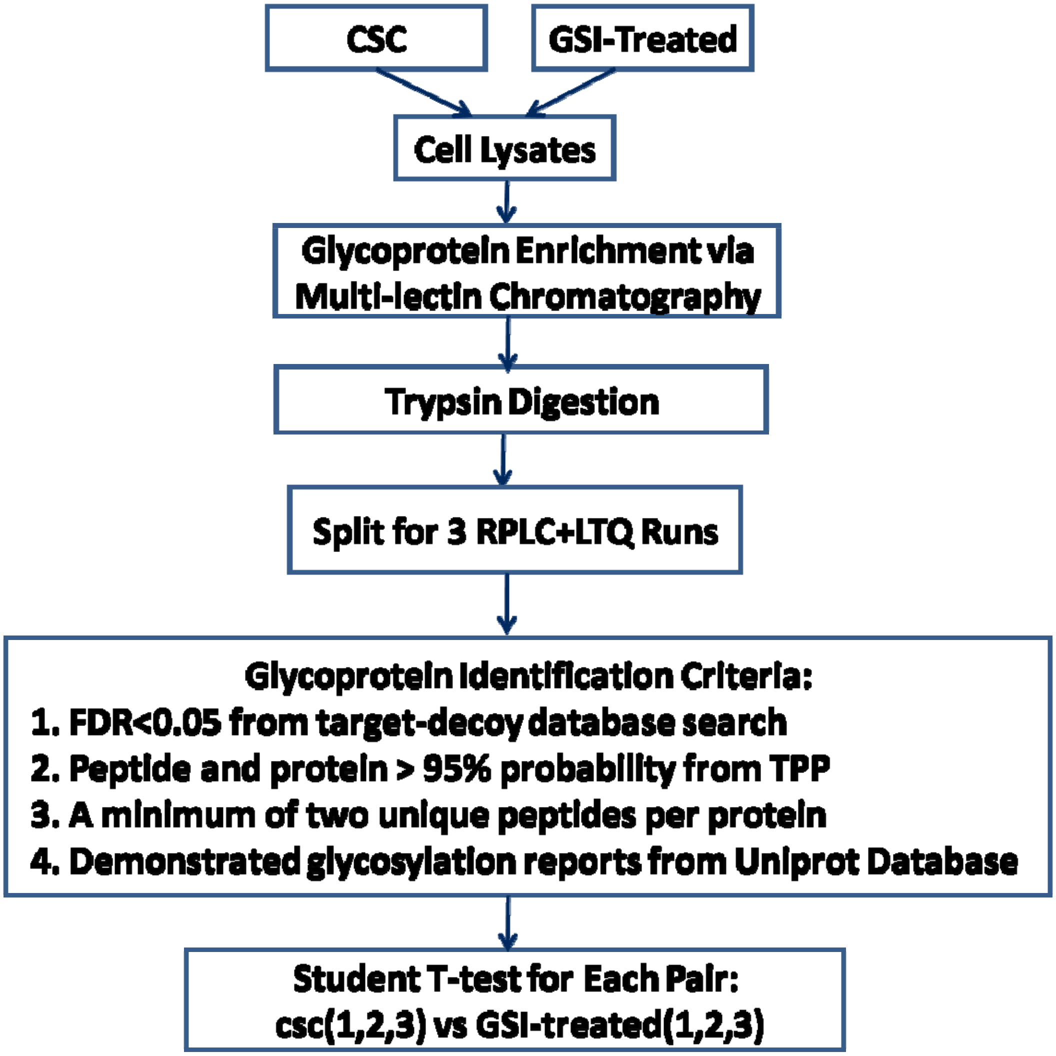
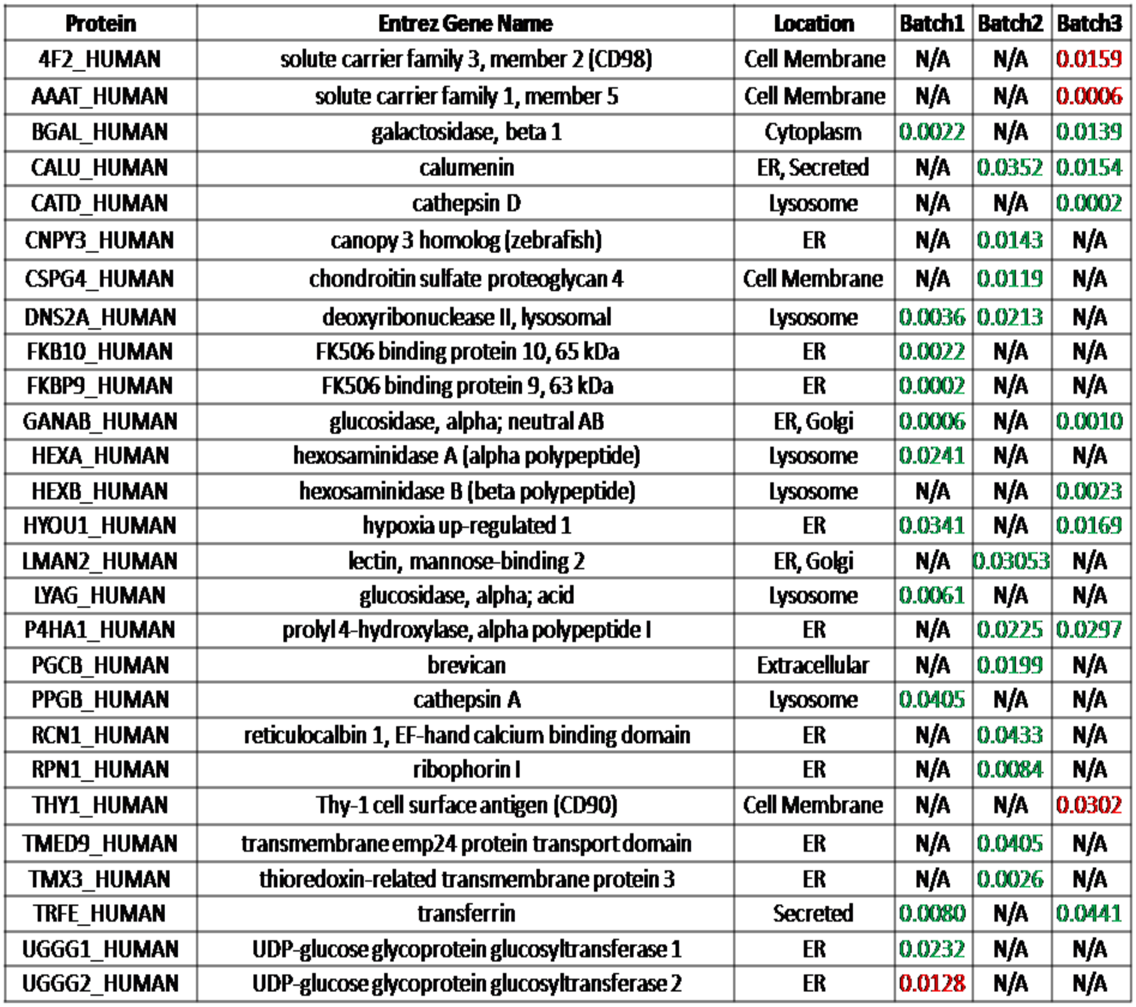

CSC
GSI-Treated
Cell Lysates
Glycoprotein Enrichment via Multi-lectin Chromatography
Trypsin Digestion
Split for 3 RPLC+LTQ Runs
Glycoprotein Identification Criteria:
1. FDR<0.05 from target-decoy database search
2. Peptide and protein > 95% probability from TPP
3. A minimum of two unique peptides per protein
4. Demonstrated glycosylation reports from Uniprot Database
Student T-test for Each Pair:
csc(1,2,3) vs GSI-treated(1,2,3)

Table 4.1: Differentially expressed glycoproteins. Three pairs of student's t-test are performed on the three biological replicates with a total of 9 technical replicates for each sample. P-values less than 0.05 for each pair of comparison are accepted and highlighted in green (down-regulation) and red (up-regulation). N/A indicates p-values larger than 0.05.



Protein	Entrez Gene Name	Location	Batch1	Batch2	Batch3
4F2_HUMAN	solute carrier family 3, member 2 (CD98)	Cell Membrane	N/A	N/A	0.0159
AAAT_HUMAN	solute carrier family 1, member 5	Cell Membrane	N/A	N/A	0.0006
BGAL_HUMAN	galactosidase, beta 1	Cytoplasm	0.0022	N/A	0.0139
CALU_HUMAN	calumenin	ER, Secreted	N/A	0.0352	0.0154
CATD_HUMAN	cathepsin D	Lysosome	N/A	N/A	0.0002
CNPY3_HUMAN	canopy 3 homolog (zebrafish)	ER	N/A	0.0143	N/A
CSPG4_HUMAN	chondroitin sulfate proteoglycan 4	Cell Membrane	N/A	0.0119	N/A
DNS2A_HUMAN	deoxyribonuclease II, lysosomal	Lysosome	0.0036	0.0213	N/A
FKB10_HUMAN	FK506 binding protein 10, 65 kDa	ER	0.0022	N/A	N/A
FKBP9_HUMAN	FK506 binding protein 9, 63 kDa	ER	0.0002	N/A	N/A
GANAB_HUMAN	glucosidase, alpha; neutral AB	ER, Golgi	0.0006	N/A	0.0010
HEXA_HUMAN	hexosaminidase A (alpha polypeptide)	Lysosome	0.0241	N/A	N/A
HEXB_HUMAN	hexosaminidase B (beta polypeptide)	Lysosome	N/A	N/A	0.0023
HYOU1_HUMAN	hypoxia up-regulated 1	ER	0.0341	N/A	0.0169
LMAN2_HUMAN	lectin, mannose-binding 2	ER, Golgi	N/A	0.03053	N/A
LYAG_HUMAN	glucosidase, alpha; acid	Lysosome	0.0061	N/A	N/A
P4HA1_HUMAN	prolyl 4-hydroxylase, alpha polypeptide I	ER	N/A	0.0225	0.0297
PGCB_HUMAN	brevican	Extracellular	N/A	0.0199	N/A
PPGB_HUMAN	cathepsin A	Lysosome	0.0405	N/A	N/A
RCN1_HUMAN	reticulocalbin 1, EF-hand calcium binding domain	ER	N/A	0.0433	N/A
RPN1_HUMAN	ribophorin I	ER	N/A	0.0084	N/A
THY1_HUMAN	Thy-1 cell surface antigen (CD90)	Cell Membrane	N/A	N/A	0.0302
TMED9_HUMAN	transmembrane emp24 protein transport domain	ER	N/A	0.0405	N/A
TMX3_HUMAN	thioredoxin-related transmembrane protein 3	ER	N/A	0.0026	N/A
TRFE_HUMAN	transferrin	Secreted	0.0080	N/A	0.0441
UGGG1_HUMAN	UDP-glucose glycoprotein glucosyltransferase 1	ER	0.0232	N/A	N/A
UGGG2_HUMAN	UDP-glucose glycoprotein glucosyltransferase 2	ER	0.0128	N/A	N/A

Figure 4.2: Comparison of protein expression level before (red bars) and after enrichment (blue bars). The horizontal axis represents spectral counts for each protein.

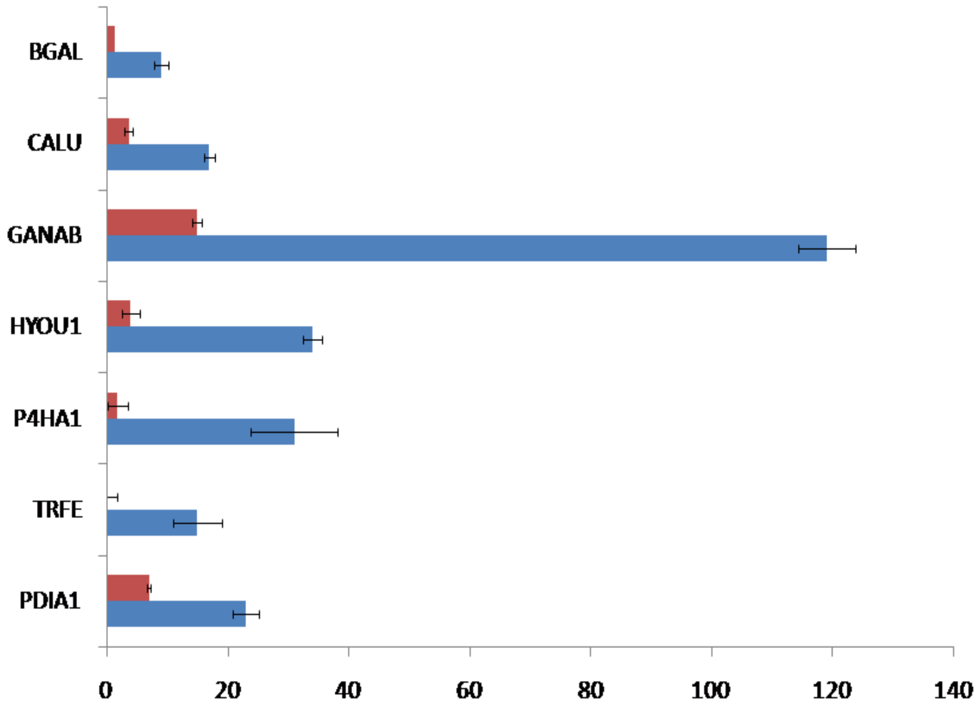
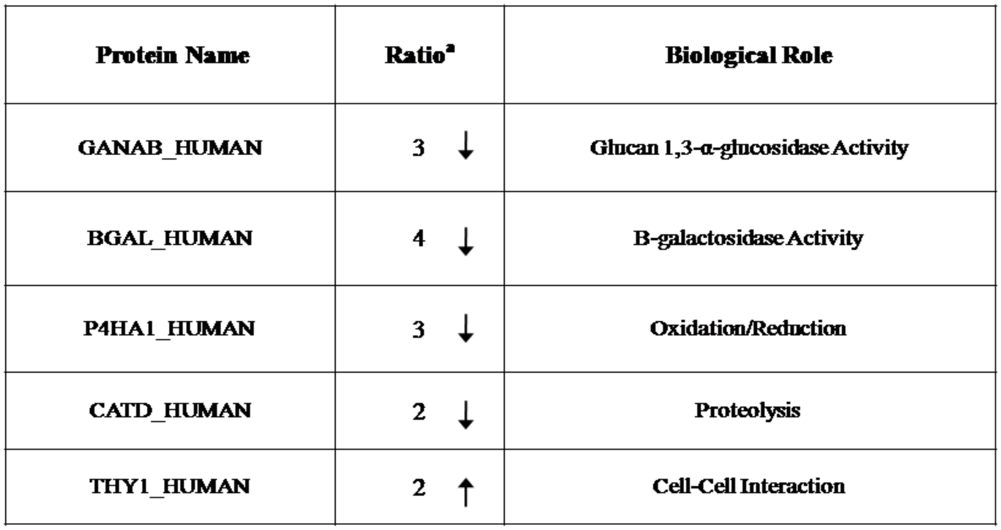
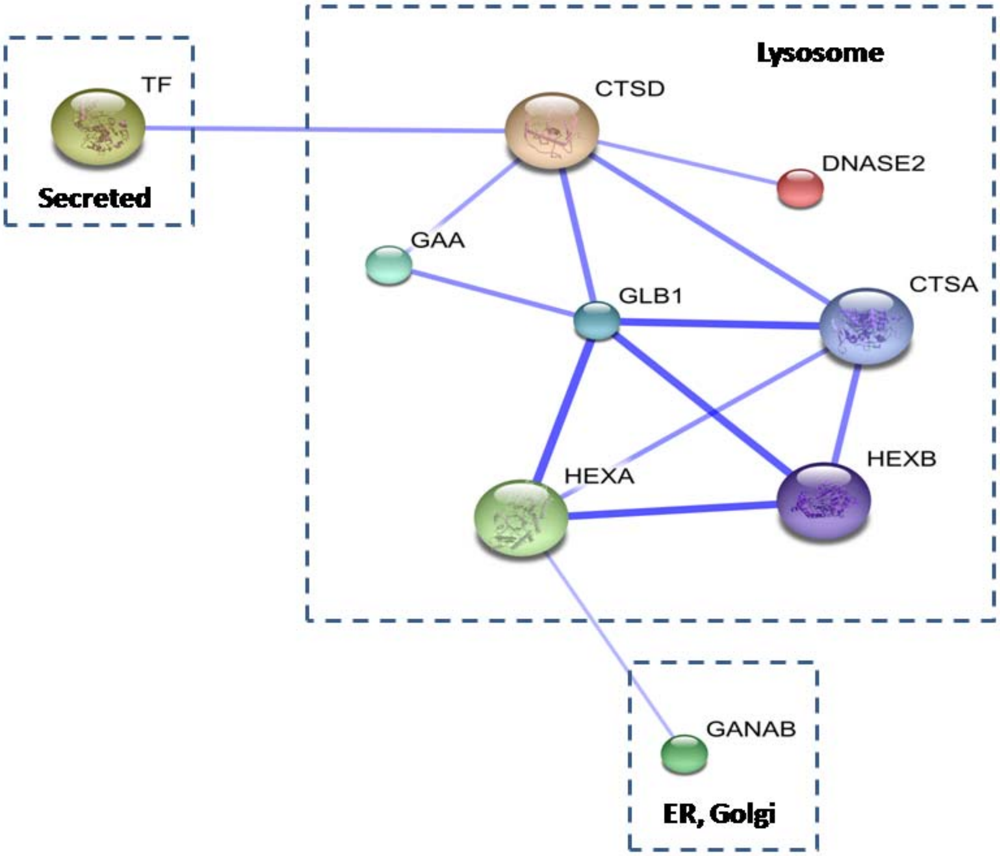


Table 4.2: Summary of the selected proteins for Western Blot. Direction of Arrows indicates up or down-regulation after GSI treatment. The ratio is calculated by averaging the fold changes from three pairs of comparisons.



Protein Name	Ratio[a]	Biological Role
GANAB_HUMAN	3 ↓	Glucan 1,3-α-glucosidase Activity
BGAL_HUMAN	4 ↓	B-galactosidase Activity
P4HA1_HUMAN	3 ↓	Oxidation/Reduction
CATD_HUMAN	2 ↓	Proteolysis
THY1_HUMAN	2 ↑	Cell-Cell Interaction

Figure 4.3: Protein-Protein interaction network generated by STRING. Each node represents a protein and each edge represents an interaction in between. A thicker line indicates stronger association.



CHAPTER 5

DOSE-DEPENDENT PROTEOMIC ANALYSIS OF GLIOBLASTOMA CANCER STEM CELLS UPON TREATMENT WITH GAMMA-SECRETASE INHIBITOR

5.1 Abstract

Notch Signaling has been demonstrated to have a central role in Glioblastoma (GBM) Cancer Stem Cells (CSCs) and we have demonstrated recently that Notch pathway blockade by γ-secretase inhibitor (GSI) depletes GBM CSCs and prevents tumor propagation both in vitro and in vivo. In order to understand the proteome alterations involved in this transformation, a dose-dependent quantitative mass spectrometry (MS) based proteomic study has been performed based on global proteome profiling and a target verification phase where both Immunoassay and a Multiple Reaction Monitoring (MRM) assay are employed. The selection of putative protein candidates for confirmation poses a challenge due to the large number of identifications from the discovery phase. A multilevel filtering strategy together with literature mining is adopted to transmit the most confident candidates along the pipeline. Our results indicate that treating GBM CSCs with GSI induces a phenotype transformation towards non-tumorigenic cells with decreased proliferation and increased differentiation, as well as elevated apoptosis. Suppressed glucose metabolism and attenuated NFR2-mediated oxidative stress response are also suggested from our data, possibly due to their crosstalk with Notch Signaling.

Overall, this quantitative proteomic based dose-dependent work complements our current understanding of the altered signaling events occurring upon the treatment of GSI in GBM CSCs.

5.2 Introduction

Glioblastoma multiforme(GBM) is the most aggressive class of brain tumors and 80% of patients with GBM survive only for 1-2 years after diagnosis[1]. The emerging evidence for the involvement of brain cancer stem cells in the initiation and propagation of brain tumors, particularly GBM, allows for the identification of more effective therapeutic targets[2]. Several groups have identified brain tumor CSCs using cell surface markers such as CD133 and CD15 [3-5], although currently there is no universally accepted collection of CSLC markers for isolation of a pure population of GBM stem cell-like cells[6]. GBM neurosphere cultures are often utilized as an alternative to provide an advanced model for investigating GBM CSCs[7].

The importance of Notch signaling in cancer has been firmly established and it is one of the most intensively studied therapeutic targets in CSCs. Increasing evidence has implicated its central role in GBM[7-10] based on its participation in regulation of self-renewal and cell fate determination in normal stem cells[11]. Therefore, the investigation of the molecular mechanism upon blocking at multiple stages of the Notch signaling cascade become essential where inhibition via γ-secretase inhibitors (GSIs) are the most utilized[6]. We have demonstrated in our previous study that Notch pathway blockade by GSI targets brain tumor CSCs through decreased proliferation and induced differentiation and apoptosis [7, 9,12].

The conventional biomarker discovery pipeline usually begins with a global unbiased screening stage which is typically MS-based. A quantitative MS proteomic approach has been demonstrated to be a powerful tool in the study of stem cells utilizing either stable isotope labeling methods or label free methods [13,15-16,21]. To gain further insight into the effects GSI exerts on Notch signaling and other potential pathways involved in GBM CSCs, we have employed a spectral counting-based label free quantitative proteomic approach to perform a large scale screening in global discovery phase. This initial profiling provides us comprehensive information about the proteome alterations which then requires verification after candidate prioritization via a multilevel filtering strategy. Also, the biomarker discovery pipeline usually involves a secondary targeted quantitative stage which traditionally relies on antibody-based protocols such as ELISA to follow up the proteomics or genomic profiling studies [17]. C

Currently there has been a trend toward the development of targeted MS as a methodology for confirmation based on the use of MRM [18, 19]. The concept of monitoring specific peptides from proteins of interest as an accurate quantification strategy is well established, because MRM offers superior sensitivity and selectivity for the targeted analytes and the precision is further increased by facilitating the chromatographic retention time as another identifier. Due to the complementarity of Immnoassay and MRM, we have explored a combination of these two assays to verify selected high-priority protein candidates. Moreover, literature mining was performed together with Ingenuity Pathway Analysis (IPA) to relate our findings to previous publications in order to broaden our current knowledge about the underlying molecular mechanisms regarding alterations occurring upon GSI treatment in GBM CSCs. A putative altered signaling network is generated to summarize our findings reflecting those in light of previous publications and those newly mined from our data.

5.3 Materials and Methods

5.3.1 Cell Culture and Treatments

GBM neurosphere cultures were maintained in Neurocult medium (Stem Cell Technologies, Vancouver, BC, Canada, http://www.stemcell.com) supplemented with epidermal growth factor (10 ng/ml) and fibroblast growth factor (10 ng/ml) as previously described[7, 20]. For treatment studies, cells were plated and allowed to grow overnight in Neurocult medium; Neurocult was then replaced the next morning with medium containing γ-secretase inhibitor([11-endo]-N-(5,6,7,8,9,10-hexahydro-6,9-methanobenzo[a][8]annulen-11-yl)-thiophene-2-sulfonamide, referred to as "GSI")[9] dissolved in dimethyl sulfoxide (DMSO) at the concentrations of 0, 2, 10, 50μM. We have shown previously that GSI can block Notch signaling pathway at Hes1 protein expression level starting at 2μM level[7,9].

5.3.2 Cell lysis and Trypsin digestion

Cells were harvested on day three and washed twice with PBS (0.01 M phosphate, 0.15 M NaCl, pH 7.4) to remove culture medium. The extraction of whole cell lysates follows the procedure as previously described[21]. Basically cell pellets were resuspended in PPS (Protein Discovery, Knoxville, TN) powder dissolved in 50 mM Ammonia Bicarbonate at a final concentration of 0.2%(m/v) together with 1% protease inhibitor cocktail. Protein concentration was determined by Micro BCA™ Protein Assay Kit (Pierce/Thermo Scientific, Rockford). Trypsin digestion, cleavage of PPS and purification of peptides were performed sequentially and also follow the same protocol[21]. Peptides were lyophilized to powder and stored in a -80 °C freezer for future use. All chemicals were purchased from Sigma unless mentioned otherwise.

5.3.3 Reversed Phase Liquid Chromatography and ESI-Ion Trap

Peptides were reconstituted in a solution of 5% ACN with 0.1% formic acid at a final concentration of 100ng/µl. Reversed phase Liquid Chromatography were performed by a Paradigm MG4 micropump system (Michrom Biosciences Inc., Auburn, CA) connected to LTQ mass spectrometer (Thermo Finnigan, San Jose, CA). Total tryptic digests of each sample (control and 3 treatments) were directly introduced into a RPLC nano column (3µm x 200 Å, 0.1 mm × 150 mm, C18 AQ particles, Michrom) after a desalting nano trap (300 × 50 mm) (Michrom). A 3hr linear gradient with 150min from 5 to 40% ACN, 15min from 5% to 80% ACN and another 15min for equilibrium to 5% ACN was used. The other LTQ parameters are the same as previously described[21]. Each sample was analyzed in triplicate with each injection of 1µg material.

5.3.4 Database Searching and Multilevle Filtering

MS/MS spectra were searched against Uniprot database by SEQUEST search engine incorporated in Proteome Discoverer (version 1.1.0.263). Searching parameters were specified as follows: (1) Fixed modification: carbamidomethylation of Cys residue with a mass shift of 57.02Da; (2) variable modification: oxidation of Met residue with a mass shift of 15.99Da; (3) two missed cleavage sites were allowed; (4) peptide ion mass tolerance: 1.4 Da; (5) fragment ion mass tolerance: 1.0 Da; (6) peptide charges +1, +2, and +3. Searching results were further uploaded to Scaffold as msf format (the default output format from Proteome Discoverer). Then a multilevel filtering strategy consisting of four checkpoints is adopted to capture the most confident identifications and differentially expressed proteins. Checkpoint-1 is based on the FDR test from target-decoy database search with a cutoff p-value less than 0.05; Checkpoint-2 is according to the Trans-Proteomic Pipeline (TPP)[22] built in Scaffold. The criteria include protein and peptide probabilities above 95% and a minimum of two unique peptides identified for each protein; Checkpoint-3 is tested by first generating three lists of differentially expressed proteins by applying the student t-test for 0 vs 2μM, 0 vs 10μM and 0 vs 50μM with a threshold p-value less than 0.05 and then retaining proteins that are present in all three lists; Checkpoint-4 is a further filtering based on literature mining to retain proteins that have been previously reported to play important roles in furthering CSLC properties and/or Tumorigenesis.

5.3.5 Western Blot

Western Blot was performed essentially as previously described[1]. Briefly, 20 μg of total proteins from each sample were separated by 4-20% SDS-PAGE and then transferred to PVDF membranes (Bio-Rad, CA). After being blocked for 2 h, the membranes were incubated with antibodies including polyclonal anti-APC5, polyclonal anti-GFAP, monoclonal anti-ENGO, monoclonal anti-PCNA, polyclonal anti-SODC and monoclonal anti-Actin (Abcam, Cambridge, CA) overnight. After washing three times, the membranes were incubated with peroxidase conjugated goat anti-rabbit or anti-mouse IgG (H+L) for 1 h. The blots were visualized with DAB stain (Vector Laboratory, WI).

5.3.6 MRM Assays

For the protein of interest, the selection criteria of proteolytic signature peptides include: (1) being identified from the LTQ analysis with high confidence; (2) a unique signature of the target protein; (2) a length of 8-20 amino acids; (3) no missed cleavage sites; (4) no post translational modifications. The MRM assay was performed with an Agilent 6410 triple quadrupole MS system equipped with an Agilent 1200 LC (Agilent Technologies, New Castle, DE) in positive ion mode. Synthetic peptides were first reconstituted in 50% methanol. Flow injection analysis (FIA) was used to optimize the fragmentor voltage and collision energy determined by the intensity of precursor ions and product ions, respectively. The two most abundant transitions from each peptide were chosen to obtain the best signal-to-noise ratio in MRM mode. C-18 column from Agilent with 1.8μm particle size and 4.6 x 50mm dimension was used for the HPLC separation. . The mobile phase is 0.1% formic acid (Solvent A) and 0.1% formic acid in acetonitrile (Solvent B). The linear gradient was 2 to 20% acetonitrile for 1.5 minutes and 20 to 95% acetonitrile for 5 minutes with a flow rate of 0.6ml/min. The desolvation gas temperature is 350 C and the capillary voltage is 4000V. The nebulizer pressure is 45 psi and the desolvation gas flow rate is 11 l/min.

For the generation of standard curves, 5, 10, 20, 40, 80, 120, 200 fmol/μl of a mixture of synthetic targeted peptides were injected and analyzed in triplicate. For the monitoring of sample response, five dose (0, 0.4, 2, 10, 50μM) points were interrogated with a total sample protein injection of 5μg and 10μg analyzed in duplicates. Data analysis was carried out by Agilent Mass Hunter Quantitative Analysis Software.

5.3.7 Ingenuity Pathway Analysis (IPA)

The three lists of differentially expressed proteins generated after checkpoint-3 were uploaded to IPA together with their fold changes as three different observations. Pathway analysis was then performed to infer the significantly altered pathways associated with GSI treatments. The significance values for analysis of pathway generation were calculated using the right-tailed Fisher's Exact Test.

5.4 Results and Discussion

5.4.1 Protein Identifications and Differential Expressions

A one-dimensional separation strategy was adopted in place of a two-dimensional approach used in our previous studies[14,21] by reducing the particle size of the nanoRPLC column from 5μm to 3μm and prolonging the gradient time from 1hr to 3hrs to avoid protein losses while also saving instrument time. A total of 1127, 929, 854 and 638 proteins were identified for each sample respectively after filtering by a threshold of FDR < 0.05. In order to improve the confidence of protein identifications, a multilevel filtering strategy is employed as shown in Figure 5.1. A total of 1707 proteins combining all replicate runs across different samples were identified after checkpoint-1 and a fraction of 672 proteins were retained after checkpoint-2. Next, three lists of differentially expressed proteins were generated for each pair (0 vs 2μM, 0 vs 10μM and 0 vs 50μM) by a cut-off p-value < 0.05, resulting in 117, 187 and 213 differentially expressed proteins respectively. Checkpoint-3 which requires the presence in all three differential expression lists further narrows down the number of putative candidates to 36.

5.4.2 Evaluation of Label-free Quantification

Reproducibility is an essential factor to evaluate the accuracy of a label-free based quantification approach. It is interrogated from two aspects in this study: variation of the number of protein identifications across three technical replicates and the correlation of spectral counts between any two technical replicates within the same sample. First, the number of total identifications passing checkpoint-1 in each replicate run is similar to each other with a coefficient of variance (CV) of 3%, 1%, 3% and 8% for each of the control and 2, 10, 50μM treatment samples, respectively. Second, high Pearson correlation coefficients are found by comparing any two replicates within the same sample. For example, the coefficient is calculated to be 0.977 between [0.973, 0.982] at 95% confidence level based on the spectral counts assigned to each identified protein between the first and the second technical replicate run of the control sample. This correlation matrix was also used to generate the clustering graph shown in Figure 5.2. The four different biological entities are clearly separated with their technical replicates grouped under the same branch. Also, the three treatment groups are more closely associated compared to the control group where 2μM and 10μM dose treatments tend to be more closely related compared to the 50μM treatment.

5.4.3 Selection of Putative Protein Candidates

The list of 36 protein candidates were further screened through literature mining to link their biological roles pertaining to CSCs properties and/or dysregulated events in tumor. A total of 15 high-priority putative candidates were selected as listed in Table 5.1 and are categorized by different functions: Proliferation, Differentiation, Apoptosis, Tumor Invasion, Oxidative Response and Glycolysis. Four proteins (PCNA, NEST, DREB, SODC) from different functional categories in this table were selected for validations by either western blot or MRM assays. Three proteins out of this table were also selected because of their association to these investigated functions: Gamma-Enolase(ENOG) (alternatively : neuron-specific enolase (NSE)) for “Proliferation”; Glial fibrillary Acidic Protein (GFAP) for Differentiation; Anaphase promoting complex5 (APC5) for Apoptosis.

5.4.4 Validation through Western Blot

The fold changes of five proteins were validated by western blot experiments for two dose treatments (2μM and 10μM) of an independent control sample. Figure 5.3 illustrates a consistent dose-dependent pattern detected between the spectral counting method and western blot experiments. PCNA is commonly used as a cell proliferation index and a good candidate for prognosis of tumor and cancer development[23]. The decreased expression pattern of PCNA detected in GBM CSCs in our study agrees with a previous report that disruption of Notch signaling by GSI in tracheal epithelial cells reduces PCNA expression[24] and our previous study showing Notch pathway inhibition by GSI reduces CSLC proliferation[7,9]. ENOG has been used as a neuron stem cell marker[63]. A previous study of different subgroups of GBM tumor-initiating cells shows 70% of ENOG positive cells have developed tumors using a mice xenograft model[25] and the expression of ENOG are only detected in high-grade GBMs[26]. Our finding of down-regulated expression of ENOG may imply that GBM CSCs exhibit a reduced tumor grade as the drug dose increases.

GFAP was chosen to verify the impact of GSI on the differentiation of GMB CSCs. It has been previously detected in ~78% differentiated brain tumor CSCs and exhibits lack of immunoreactivity in undifferentiated CSCs[27]. Our result of up-regulated GFAP expression indicates the treatment of GSI drives GBM CSCs towards a more differentiated state and the differentiation degree is positively correlated with the drug dosage. For the verification of altered apoptotic activity upon GSI treatment, Anaphase-Promoting Complex, Subunit 5 (APC5) which is a subunit of the multiprotein complex that controls mitotic progression[28] was tested here. It is hypothesized that the abnormal regulation of APC may be involved in malignant transformation through chromosome instability[29] and the inhibition may lead to cell death[30]. Thus, our observation of reduced expression level of APC5 may be an indication of cell cycle failure in GBM CSCs after treatment with GSI which in turn promotes cell death. SODC is an important anti-oxidant enzyme which protects cells from free radical attack and it has been demonstrated to play a critical role in "Reactive Oxygen Species" (ROS) defense and is associated with chemoresistance and malignancy grade in astrocytic brain tumors[31]. Thus, the reduced expression level of SODC detected in our experiments may imply that the GSI treatment may render GBM CSCs more vulnerable to apoptosis by attenuating their cell defense system. Overall, the western blot results of these five proteins suggest three areas of impact on GBM CSCs upon GSI treatment: reduced proliferative potential, increased differentiation, and enhanced apoptotic activity. All of these could be viewed as a decrease of stem cell properties, leading to a phenotype change.

5.4.5 Validation through MRM

In addition to immunoassay, MRM is also employed as another orthogonal verification strategy to validate the turnover of two important candidates in this study. NESTIN has been identified as a neural stem cell marker and a GBM CSLC marker[9, 27, 32]. We and others have demonstrated previously that NESTIN expression is enhanced by Notch signaling in medulloblastoma[9] and GBM[33] and is inhibited by GSI in a dose-dependent fashion[7] in GBM CSCs. These findings were obtained by immunostaining related approaches. Herein, a MS-based method was used to confirm the impact of the drug on NESTIN expression in GBM CSCs. Another protein candidate chosen for MRM is Developmentally-regulated Brain Protein (DREB/Drebrin) which functions in cell migration, extension of neuronal processes via binding to F-actin and regulates neuronal actin dynamics and plasticity[34, 35]. A new role of Drebrin has just been recently uncovered that this protein exerts as an important modulator of the chemokine receptor CXCR4 and the knockdown of Drebrin impairs CXCR4 function[36].

The linearity of the targeted response was evaluated by synthetic peptides. Two peptides from NESTIN and one peptide from Drebrin were chosen according to the peptide selection criteria described in the experimental section. For each targeted peptide, the two most intense and stable MS/MS fragment ions were selected for the generation of two transitions as listed in Table 5.2. Figure 5.4 shows an extracted MRM ion chromatogram for all six transitions. Although Drebrin peptide (precursor ion m/z: 717.1) and NESTIN peptide (precursor ion m/z: 691.6) were eluted with a slight retention time difference of 0.84 second, the response from the mass analysis shows good linearity in the standard curve with an average of $R^2 = 0.958$ and $R^2 = 0.968$ from two transitions for each peptide, respectively, ranging from 10fmol/µl to 200fmol/µl. The other NESTIN peptide (precursor ion m/z: 631.5) exhibits superior linearity with an average of $R^2 = 0.996$ when monitored within the same concentration range. A narrower range of standard curve was further generated for an improved quantification after the examination of sample response.

The synthetic peptides used as external calibration standards in our MRM experiments were not isotope- labeled and hence they were not spiked into the sample to avoid interference from identical response, where the main focus of this study is to investigate the change of protein expression as a function of drug dosage. Thus, absolute quantification is not needed and it is sufficient to use external calibration curves to calculate the relative response of sample analytes. Also, the overall goal is to verify if the fold change results obtained from the MRM method correlate with those from label free method. In the latter case, spectral counts are used as surrogate measurements and essentially they represent a relative quantification as well. Therefore, the results obtained from our MRM assays with external calibrations provide adequate information comparable to the label-free results after converting the response from each dosage into a ratio over the control.

The extracted MRM responses from 10μg total injections have generally higher signal-to-noise ratio than 5μg total injections. This is especially true at 10μM and 50μM dose points where the amount of targeted peptides is approaching the lowest limit of detection. The dose-dependent results from label free data and the MRM data are shown in Figure 5.5. Basically, the trends of the fold change detected from these two methods are similar to each other: both NESTIN and Drebrin exhibit reduced expression level as a function of increased GSI dosage. The correlation between these two methods is superior for NESTIN. The expression level is reduced by ~1.5 to 2.5 fold when increasing the dose from 0.4μM to 10μM, while no difference can be observed between 10μM to 50μM from both label free and MRM results. This is also in line with our previous report that a reduced mRNA NESTIN expression is detected from ~1.5 to 4 fold after treating GBM CSCs with GSI at 2, 10 and 50μM[7]. The overall dose-dependent pattern for Drebrin between these two methods correlates as well, although the reduced fold changes detected from MRM data indicates a ~1.5 to 2-fold change whereas the label free data indicates a ~2.5 to 7.5-fold change. The correlation for NESTIN is better than Drebrin as shown in the lower panel of Figure 5.5. Generally, in the low spectral counts range (<10), label free data tend to over-estimate the differential expression due to the lowest limit of detection cutoff. For example, Drebrin is assigned a mean spectral count value of 1 at 10µM and 50µM, indicating it has already reached the lowest cutoff. As a comparison, the mean of the lowest spectral counts for NESTIN is 18 and the CV is also lower than that of Drebrin. Both of them indicate that the measurement for NESTIN is more accurate when using the same label free quantification method. In addition, the small difference detected for the fold changes of Drebrin between label free and the MRM method largely lies in the limitation of data-dependent acquisition strategy employed by the LTQ.

5.4.6 Altered Signaling Events upon GSI Treatment

Based on the knowledge obtained from a combination of literature mining and data mining, candidate proteins listed in Table 5.1 and three proteins out of this table that have been verified (APC5, GFAP and ENOG) were integrated to construct a putative altered signaling network after treating GBM CSCs with GSI, as depicted in Figure 5.6. In addition, another three proteins were also imported**:** Thioredoxin (THIO/Trx), T-complex protein 1 subunit eta (TCPH/CCT7) and Hexokinase-1(HKX1). The first two proteins successfully passed checkpoint-2 with a p-value less than 0.05 in 0 vs 10µM and 0 vs 50µM treatments although they are not differentially expressed after 2µM treatment. The third protein passed checkpoint-1 but did not pass checkpoint-2 and it shows decreased expression upon 10 and 50µM treatment. The correlation between these proteins and their targeted functions are listed in Table 5.3.

The majority of altered signaling events depicted in this figure are mediated through the Notch signaling cascade. Upon blockade of the Notch pathway, proliferation of CSCs is selectively reduced[7] which is supported by our finding of decreased expression level of four key proteins: NESTIN, ENOG, PCNA and PA2G4. Also, differentiation is induced[7] which can be indicated by the elevated expression levels of two marker proteins for differentiated neural cells: GFAP and TBB3. Previous studies in Multiple myeloma report that activation of Notch signaling inhibits apoptosis while inhibition of Notch induces apoptosis [37, 38]. We also have demonstrated that Notch inhibition by GSI induces apoptosis in medulloblastoma and GBM[7,9,12]. Our results are in line with this where the blockade of Notch activates apoptosis in GBM CSCs. This is supported by the detection of a reduced expression level of APC5, PDIA1 VDAC1, TCPB and 1433 proteins, all of which have been reported to be negatively correlated with apoptosis [30, 39-46]. In addition, decreased tumor invasion capability is inferred and supported by the detection of reduced expression of DREB and MYH9 which are identified to be positively correlated with metastasis[36, 47, 48].

Another signaling cascade we speculate to be down-regulated is NFR2-mediated Oxidative Response which contributes to cellular protection against oxidative insults and chemical carcinogens[49]. The decreased expression levels of its downstream transcriptional gene products, such as anti-oxidant proteins, SODC and THIO and a molecular chaperone protein, TCPH, imply that this cellular defense system against "ROS" has been attenuated[31, 50-52]. As aforementioned, treating GBM CSCs with GSI tends to abrogate the stem cell properties[7]. Taking these together, it is reasonable to infer that the decreased oxidative defense capability is also correlated with a phenotype transformation from CSCs towards non-tumorigenic cells. This hypothesis is in light of the recent significant discoveries that subsets of CSCs in some tumors contain an enhanced defense system compared to non-tumorigenic progeny suggested by lower “ROS” levels[53, 54]. However, it is uncertain at this point that the down-regulation of NFR2-mediated oxidative response is attributed to the direct impact from the impaired γ-secretase activity or via its newly proposed crosstalk mechanism with Notch signaling[55].

In addition, the glucose metabolism pathway Glycolysis is also suggested to be down-regulated. Most cancer cells rely on anaerobic metabolism even with plenty of oxygen other than mitochondrial oxidative phosphorylation for normal differentiated cells and the glycolytic rate is increased to compensate for the less efficient production of ATP**,** a phenomenon referred to as the “Warburg effect”[56]. The expression of two out of three rate-limiting key enzymes in Glycolysis: KPYM and HKX1, were found to be decreased from our label free quantitative data, which may imply a decrease of glycolytic rate after blockade of Notch signaling**.** This inference regarding the relationship between Notch pathway and Glycolysis could be supported by previous investigations showing that Notch signals promote glucose metabolism mediated by the PI3K/AKT pathway[57] and our previous finding showing that Notch pathway blockade by GSI reduces AKT phoshorylation in medulloblastoma and GBM[7,9,12]. Another study of Pre-T cells has also shown that the withdrawal of Notch signaling reduced AKT phosphorylation and decreased glycolytic rate[58]. Moreover, the PI3K/AKT pathway has been shown to stimulate aerobic glycolysis in cancer cells[59] and directly enhance glucose capture by HKX1[60]. Therefore, we hypothesize that the blockade of Notch

Pathway upon GSI treatment decreases PI3K/AKT signaling which further suppresses Glycolysis. Another explanation to the hypothetical decreased glucose metabolism is that GSI treatment may induce a mechanistic switch back to aerobic metabolism so that the accelerated production of pyruvate is not needed; hence the glucose metabolism is down-regulated. There is no available evidence indicating that the impaired γ-secretase activity has a direct impact on the suppression of Glycolysis.

Another interesting link between NFR2 oxidative response and Glycolysis is the level of "ROS". The generation of "ROS" has been postulated to be increased as the glycolytic rate is reduced[60] and also as the oxidative defense decreased in non-tumorigenic cells compared to CSCs[54]. We have discussed above our inferences that the treatment of GSI drives a transformation towards non-tumorigenic cells, suppresses the cell defense capability and down-regulates aerobic glycolysis. Thus, it is also reasonable to infer that the level of "ROS" is increased after the treatment in GBM CSCs, although it is uncertain about the alteration mechanism of mitochondrial oxidative phosphorylation which is the major cellular source of "ROS" production.

5.4.7 Ingenuity Pathway Analysis

To gain additional insight from our data, an alternative data mining tool was utilized to construct significantly affected canonical pathways upon GSI treatment by IPA. Glycolysis, NFR2-mediated Oxidative Stress Response and PI3K/AKT signaling are captured as the most significant canonical pathways which provide another piece of evidence to support our hypothesis from a bioinformatics perspective. Other important signaling pathways are also shown including VEGF signaling, Cell Cycle and Hypoxia Signaling etc. This may be attributed to the demonstrated crosstalk between Notch signaling and these pathways[61-62] or GSI may also have direct impact on these pathways. One needs to be careful about the interpretation of the level of the significance between the treatments. The length of the bar indicates a level of association that is by no means indicative of either up or down regulation. Also, most of the pathways appear to be increasingly significant in the third treatment (50µM GSI). This may be because more proteins are differentially expressed between the treatment and the control as the dosage increases. Thus, the increased number of imported proteins may affect the outcome of the statistic algorithm (Fisher's Exact T-test) adopted by IPA for the calculation of significance by inducing a smaller p-value.

5.5 Conclusion

In summary, this work adopts a label free quantitative global proteomic approach together with Immunoassay and MRM assays to conduct a dose-dependent investigation on the proteome alterations upon the treatment of GSI in GBM CSCs. It demonstrates a work-flow from global discovery, candidate prioritization to verification phase which could be applied to other studies as well. By coupling our results with previous literature reports from us and others, a putative signaling network consisting of 21 candidate proteins with 7 being verified is generated to reflect our inference of the underlying molecular alterations upon GSI treatment. The downstream effects resulting from the blockade of Notch signaling are suggested to include a reduced proliferative potential, an increased differentiation and an elevated apoptotic activity, leading to a phenotype transformation towards non-tumorigenic cells. Novel involvement of the down-regulated NFR2-mediated oxidative stress response and Glycolysis are implied as a consequence of

GSI treatment, possible due to their crosstalk to Notch signaling. These findings regarding the alterations occurred on the proteome level and the signaling/metabolic pathway level provide enriched information that could broaden our current knowledge about drug mechanism, contributing to the identification of novel drug targets to develop better therapies for treating this dismal disease.

5.6 References

1. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B. *et al.*, *Radiotherapy plus concomitant adjuvant temozolomide for glioblastoma.* N Engl J Med, 2005. 352(10): p.987-96
2. Vescovi AL, G.R., Reynolds BA., *Brain tumour stem cells.* Nat Rev Cancer, 2006. **6**(6): p. 425-36.
3. Read TA, F.M., Markant SL, McLendon RE, *et al.*, *Identification of CD15 as a marker for tumor-propagating cells in a mouse model of medulloblastoma.* Cancer Cell, 2009. **15**(2): p. 135-47.
4. Ward RJ, L.L., Graham K, Satkunendran T, *et al.*, *Multipotent CD15+ cancer stem cells in patched-1-deficient mouse medulloblastoma.* Cancer Res, 2009. **69**(11): p. 4682-90.
5. Son MJ, W.K., Nam DH, Lee J, Fine HA., *SSEA-1 is an enrichment marker for tumor-initiating cells in human glioblastoma.* Cell Stem Cell, 2009. **4**(5): p. 440-52.
6. Gilbert CA, R.A., *Cancer stem cells: cell culture, markers, and targets for new therapies.* J Cell Biochem, 2009. **108**(5): p. 1031-8.
7. Fan X, K.L., Zhu TS, Soules ME, *et al.*, *NOTCH pathway blockade depletes CD133-positive glioblastoma cells and inhibits growth of tumor neurospheres and xenografts.* Stem Cells, 2010. **28**(1): p. 5-16.
8. Stockhausen MT, K.K., Poulsen HS., *The functional role of Notch signaling in human gliomas.* Neuro Oncol, 2010. **12**(2): p. 199-211.
9. Fan X, M.W., Khaki L, Stearns D, *et al.*, *Notch pathway inhibition depletes stem-like cells and blocks engraftment in embryonal brain tumors.* Cancer Res, 2006. **66**(15): p. 7445-52.
10. Wang J, W.T., Lathia JD, Hjelmeland AB, *et al.*, *Notch promotes radioresistance of glioma stem cells.* Stem Cells, 2010. **28**(1): p. 17-28.
11. S., C., *Notch signaling in stem cell systems.* Stem Cells, 2006. **24**(11): p. 2437-47.
12. Fan X, E.C., *Medulloblastoma stem cells.* J Clin Oncol, 2008. **26**(17): p. 2821-7.
13. Steiniger SC, C.J., Krüger JA, Yates J 3rd, *et al.*, *Quantitative mass spectrometry identifies drug targets in cancer stem cell-containing side population.* Stem Cells, 2008. **26**(12): p. 3037-46.

14. Dai L, L.C., Shedden KA, Misek DE, et al., *Comparative proteomic study of two closely related ovarian endometrioid adenocarcinoma cell lines using cIEF fractionation and pathway analysis.* Electrophoresis, 2009. **30**(7): p. 1119-31.

15. Bendall SC, H.C., Campbell JL, Stewart MH, *et al.*, *An enhanced mass spectrometry approach reveals human embryonic stem cell growth factors in culture.* Mol Cell Proteomics, 2009. **8**(3): p. 421-32.

16. Williamson AJ, S.D., Blinco D, Unwin RD, *et al.*, *Quantitative proteomics analysis demonstrates post-transcriptional regulation of embryonic stem cell differentiation to hematopoiesis.* Mol Cell Proteomics, 2008. **7**(3): p. 459-72.

17. Kitteringham NR, J.R., Lane CS, Elliott VL, *et al.*, *Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics.* J Chromatogr B Analyt Technol Biomed Life Sci, 2009. **877**(13): p. 1229-39.

18. Prakash A, T.D., Frewen B, Maclean B, *et al.*, *Expediting the Development of Targeted SRM Assays: Using Data from Shotgun Proteomics to Automate Method Development.* J Proteome Res, 2009. **8**(6): p. 2733-9.

19. Lopez MF, K.R., Sarracino DA, Prakash A, *et al.*, *Mass Spectrometric Discovery and Selective Reaction Monitoring (SRM) of Putative Protein Biomarker Candidates in First Trimester Trisomy 21 Maternal Serum.* J Proteome Res, 2010.

20. Galli R, B.E., Orfanelli U, Cipelletti B, *et al.*, *Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma.* Cancer Res, 2004. **64**(19): p. 7011-21.

21. Dai L, L.C., Shedden KA, Lee CJ, *et al.*, *Quantitative Proteomic Profiling Studies of Pancreatic Cancer Stem Cells.* J Proteome Res, 2010. **9**(7): p. 3394-402.

22. Keller A, N.A., Kolker E, Aebersold R., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**(20): p. 5383-92.

23. J Liu, N.J., H Hong, T Luo, *et al.* , *Proliferating cell nuclear antigen (PCNA) as a marker of cell proliferation in the marine dinoflagellate Prorocentrum donghaiense Lu and the green alga Dunaliella salina Teodoresco.* Journal of Applied Phycology, 2005. **17**: p. 323-30.

24. Ma XB, J.X., Liu YL, Wang LL, *et al.*, *Expression and role of Notch signalling in the regeneration of rat tracheal epithelium.* Cell Prolif, 2009. **42**(1): p. 15-28.

25. Prestegarden L, S.A., Wang J, Sleire L, Skaftnesmo KO, *et al.*, *Glioma cell populations grouped by different cell type markers drive brain tumor growth.* Cancer Res, 2010. **70**(11): p. 4274-9.

26. Rebetz J, T.D., Persson A, Widegren B, *et al.*, *Glial progenitor-like phenotype in low-grade glioma and enhanced CD133-expression and neuronal lineage differentiation potential in high-grade glioma.* PLoS One, 2008. **3**(4): p. e1936.

27. Singh SK, C.I., Terasaki M, Bonn VE, *et al.*, *Identification of a cancer stem cell in human brain tumors.* Cancer Res, 2003. **63**(18): p. 5821-8.

28. Park KH, C.S., Eom M, Kang Y., *Downregulation of the anaphase-promoting complex (APC)7 in invasive ductal carcinomas of the breast and its clinicopathologic relationships.* Breast Cancer Res, 2005. **7**(2): p. R238-47.

29. Bentley AM, W.B., Goldberg ML, Andres AJ., *Phenotypic characterization of Drosophila ida mutants: defining the role of APC5 in cell cycle progression.* J Cell Sci, 2002. **115**(pt5): p. 949-61.

30. Wang Q, M.-L.C., Couzon F, Surbiguet-Clippe C, *et al.*, *Alterations of anaphase-promoting complex genes in human colon cancer cells.* Oncogene, 2003. **22**(10): p. 1486-90.

31. Haapasalo H, K.M., Paunul N, Kinnula VL, *et al.*, *Expression of antioxidant enzymes in astrocytic brain tumors.* Brain Pathol, 2003. **13**(2): p. 155-64.

32. Hemmati HD, N.I., Lazareff JA, Masterman-Smith M, *et al.*, *2003.* Proc Natl Acad Sci U S A, Cancerous stem cells can arise from pediatric brain tumors. **100**(25): p. 15178-83.

33. Shih AH, H.E., *Notch signaling enhances nestin expression in gliomas.* Neoplasia, 2006. **8**(12): p. 1072-82.

34. Sekino Y, K.N., Shirao T., *Role of actin cytoskeleton in dendritic spine morphogenesis.* Neurochem Int, 2007. **51**(2-4): p. 92-104.

35. Shiraishi-Yamaguchi Y, S.Y., Sakai R, Mizutani A, Knöpfel T, *et al.*, *Interaction of Cupidin/Homer2 with two actin cytoskeletal regulators, Cdc42 small GTPase and Drebrin, in dendritic spines.* BMC Neurosci, 2009. **10**: p. 25-38.

36. Pérez-Martínez M, G.-A.M., Cabrero JR, Barrero-Villar M, *et al.*, *F-actin-binding protein drebrin regulates CXCR4 recruitment to the immune synapse.* J Cell Sci, 2010. **123**(pt 7): p. 1160-70.

37. Nefedova Y, S.D., Bolick SC, Dalton WS, *et al.*, *Inhibition of Notch signaling induces apoptosis of myeloma cells and enhances sensitivity to chemotherapy.* Blood, 2008. **111**(4): p. 2220-9.

38. Jia XX, L.Z., Wang H, *Suppressive effect of Notch signal activation on apoptosis of multiple myeloma cells.* Zhongguo Shi Yan Xue Ye Xue Za Zhi, 2004. **12**: p. 335-339.

39. Lovat PE, C.M., Armstrong JL, Martin S, Pagliarini V, *et al.*, *Increasing melanoma cell death using inhibitors of protein disulfide isomerases to abrogate survival responses to endoplasmic reticulum stress.* Cancer Res, 2008. **68**(13): p. 5363-9.

40. Cui Y, Z.H., Zhu Y, Guo X, *et al.*, *Proteomic analysis of testis biopsies in men treated with injectable testosterone undecanoate alone or in combination with oral levonorgestrel as potential male contraceptive.* J Proteome Res, 2008. **7**(9): p. 3984-93.

41. Pathil A, A.S., Venturelli S, Mascagni P, *et al.*, *HDAC inhibitor treatment of hepatoma cells induces both TRAIL-independent apoptosis and restoration of sensitivity to TRAIL.* Hepatology, 2006. **43**(3): p. 425-34.

42. Singh TR, S.S., Srivastava RK., *HDAC inhibitors enhance the apoptosis-inducing potential of TRAIL in breast carcinoma.* Oncogene, 2005. **24**(29): p. 4609-23.

43. Zhao ZL, L.Q., Zheng YB, Chen LY, *et al.*, *The aberrant expressions of nuclear matrix proteins during the apoptosis of human osteosarcoma cells.* Anat Rec (Hoboken), 2010. **293**(5): p. 813-20.

44. Xing H, Z.S., Weinheimer C, Kovacs A, *et al.*, *14-3-3 proteins block apoptosis and differentially regulate MAPK cascades.* EMBO J, 2000. **19**(3): p. 349-58.

45. M., R., *14-3-3 proteins in apoptosis.* Braz J Med Biol Res, 2003. **36**(4): p. 403-8.

46. Cao W, Y.X., Zhou J, Teng Z, *et al.*, *Targeting 14-3-3 protein, difopein induces apoptosis of human glioma cells and suppresses tumor growth in mice.* Apoptosis, 2010. **15**(2): p. 230-41.

47. V., B., *Myosin II motor proteins with different functions determine the fate of lamellipodia extension during cell spreading.* PLoS One, 2010. **5**(1): p. e8560.
48. Medjkane S, P.-S.C., Gaggioli C, Sahai E, Treisman R., *Myocardin-related transcription factors and SRF are required for cytoskeletal dynamics and experimental metastasis.* Nat Cell Biol, 2009. **11**(3): p. 257-68.
49. Cho HY, R.S., Debiase A, Yamamoto M, Kleeberger SR., *Gene expression profiling of NRF2-mediated protection against oxidative injury.* Free Radic Biol Med, 2005. **38**(3): p. 325-43.
50. Gómez-Pastor R, P.-T.R., Cabiscol E, Ros J, Matallana E., *Reduction of oxidative cellular damage by overexpression of the thioredoxin TRX2 gene improves yield and quality of wine yeast dry active biomass.* Microb Cell Fact, 2010. **9**: p. 9-22.
51. Ronkainen H, V.M., Kauppila S, Soini Y, *et al.*, *Increased BTB-Kelch type substrate adaptor protein immunoreactivity associates with advanced stage and poor differentiation in renal cell carcinoma.* Oncol Rep, 2009. **21**(6): p. 1519-23.
52. Go YM, C.S., Orr M, Gernert KM, Jones DP., *Gene and protein responses of human monocytes to extracellular cysteine redox potential.* Toxicol Sci, 2009. **112**(2): p. 354-62.
53. Kai K, A.Y., Kamiya T, Saya H., *Breast cancer stem cells.* Breast Cancer Res, 2010. **17**(2): p. 80-5.
54. Diehn M, C.R., Lobo NA, Kalisky T, *et al.*, *Association of reactive oxygen species levels and radioresistance in cancer stem cells.* Nature, 2009. **458**(7239): p. 780-3.
55. Wakabayashi N, S.S., Slocum SL, Agoston ES, *et al.*, *Regulation of notch1 signaling by nrf2: implications for tissue regeneration.* Sci Signal, 2010. **3**(130): p. ra52.
56. O., W., *On the origin of cancer cells.* Science, 1956. **123**(3191): p. 309-14.
57. Palomero T, Dominguez M, Ferrando AA., *The role of the PTEN/AKT Pathway in NOTCH1-induced leukemia*. Cell Cycle, 2008. **7**(8): p. 965-70.
58. Ciofani M, Z.-P.J., *Notch promotes survival of pre-T cells at the beta-selection checkpoint by regulating cellular metabolism.* Nat Immunol, 2005. **6**(9): p. 881-8.
59. Elstrom RL, B.D., Buzzai M, Karnauskas R, *et al.*, *Akt stimulates aerobic glycolysis in cancer cells.* Cancer Res, 2004. **64**(11): p. 3892-9.
60. Vander Heiden MG, C.L., Thompson CB., *Understanding the Warburg effect: the metabolic requirements of cell proliferation.* Science, 2009. **324**(5930): p. 1029-33.
61. Kanamori M, K.T., Nigro JM, Feuerstein BG, *et al.*, *Contribution of Notch signaling activation to human glioblastoma multiforme.* J Neurosurg, 2007. 106(3): p. 417-27.
62. Lee SH, K.M., Han HJ., *Arachidonic acid potentiates hypoxia-induced VEGF expression in mouse embryonic stem cells: involvement of Notch, Wnt, and HIF-1alpha.* Am J Physiol Cell Physiol, 2009. **297**(1): p. C207-16.
63. Mitchell KE, W.M., Mitchell BM, Martin P, *et al.*, *Matrix cells from Wharton's jelly form neurons and glia.* Stem Cells, 2003. **21**(1): p. 50-60.

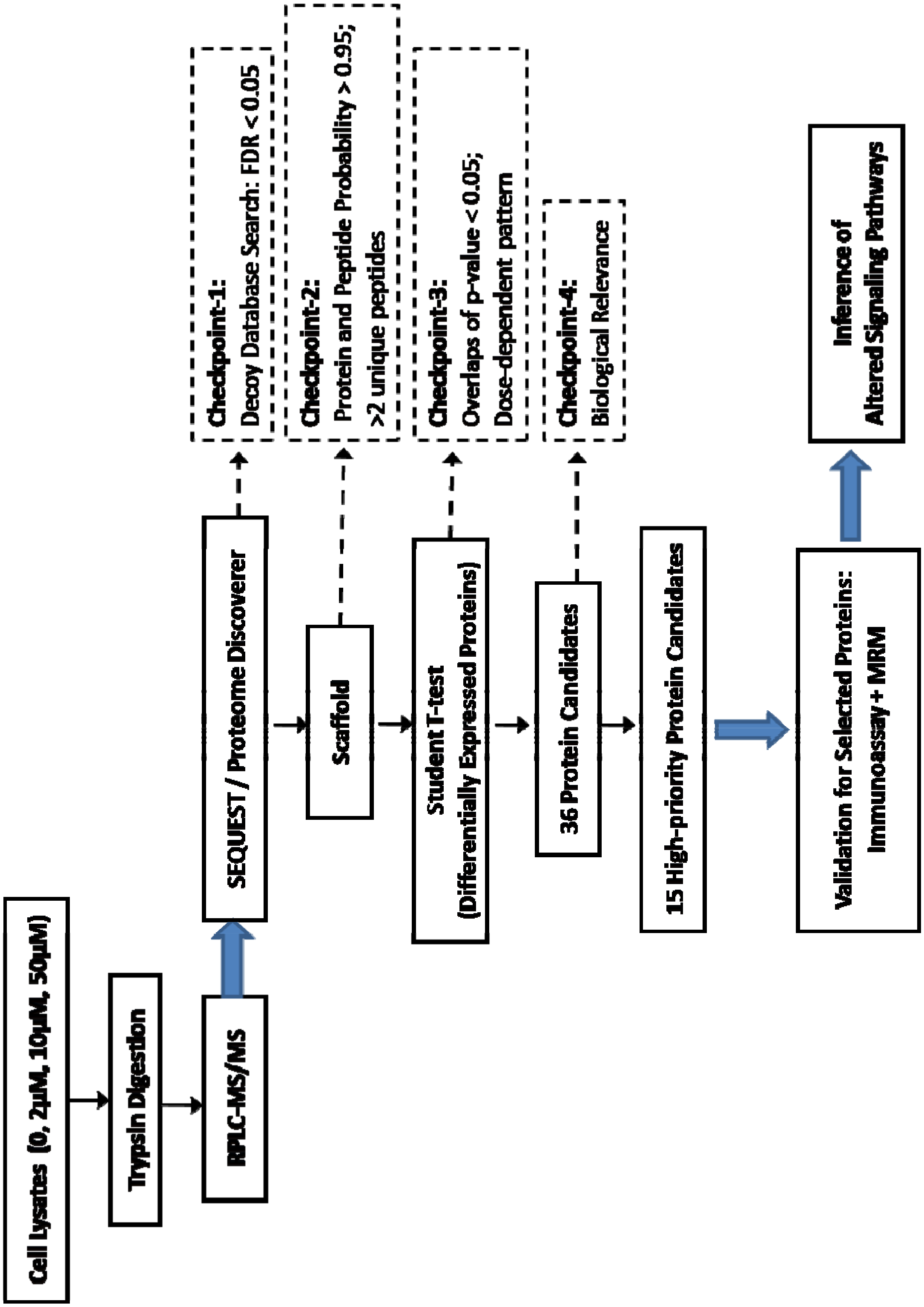

Cell Lysates (0, 2μM, 10μM, 50μM)
Trypsin Digestion
RPLC-MS/MS
SEQUEST / Proteome Discoverer
Scaffold
Student T-test
(Differentially Expressed Proteins)
36 Protein Candidates
15 High-priority Protein Candidates
Validation for Selected Proteins:
Immunoassay + MRM
Inference of
Altered Signaling Pathways
Checkpoint-1:
Decoy Database Search: FDR < 0.05
Checkpoint-2:
Protein and Peptide Probability > 0.95;
>2 unique peptides
Checkpoint-3:
Overlaps of p-value < 0.05;
Dose-dependent pattern
Checkpoint-4:
Biological Relevance

Figure 5.1: Overall Workflow

Figure 5.2: Cluster Analysis based on the correlation matrix. Pearson correlation coefficients are shown between technical replicates within the same group.

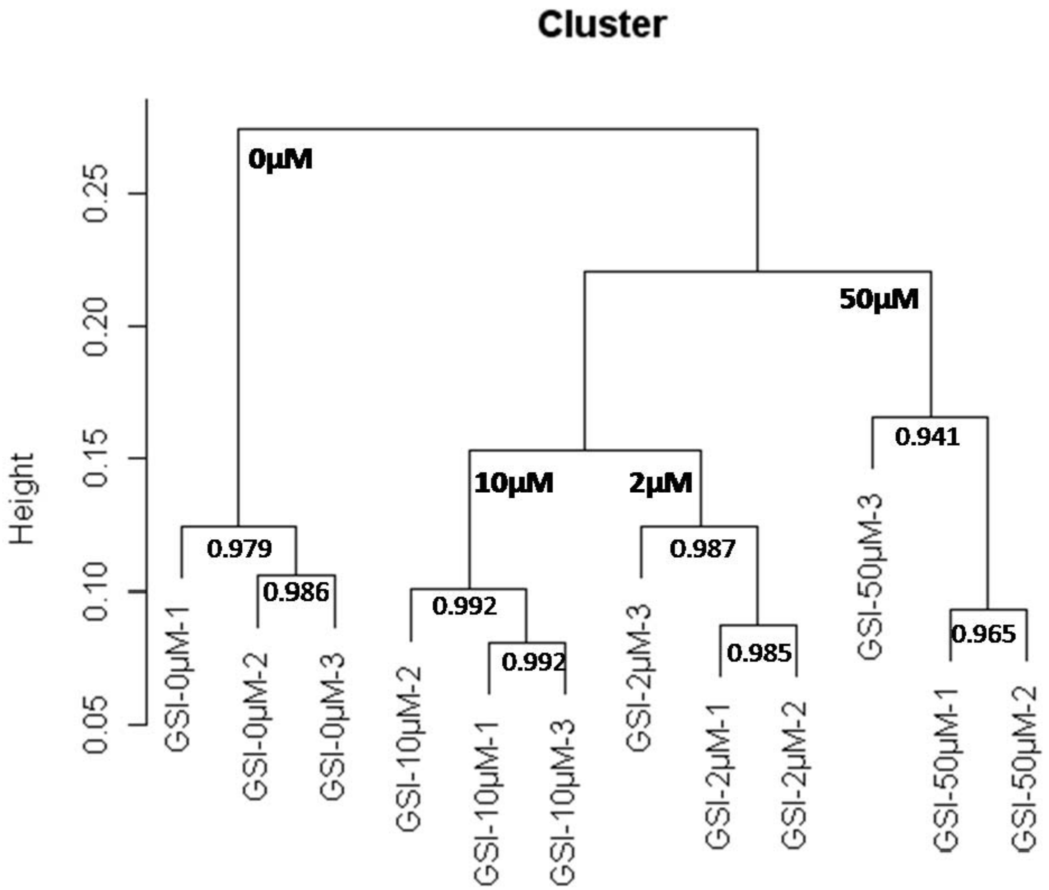
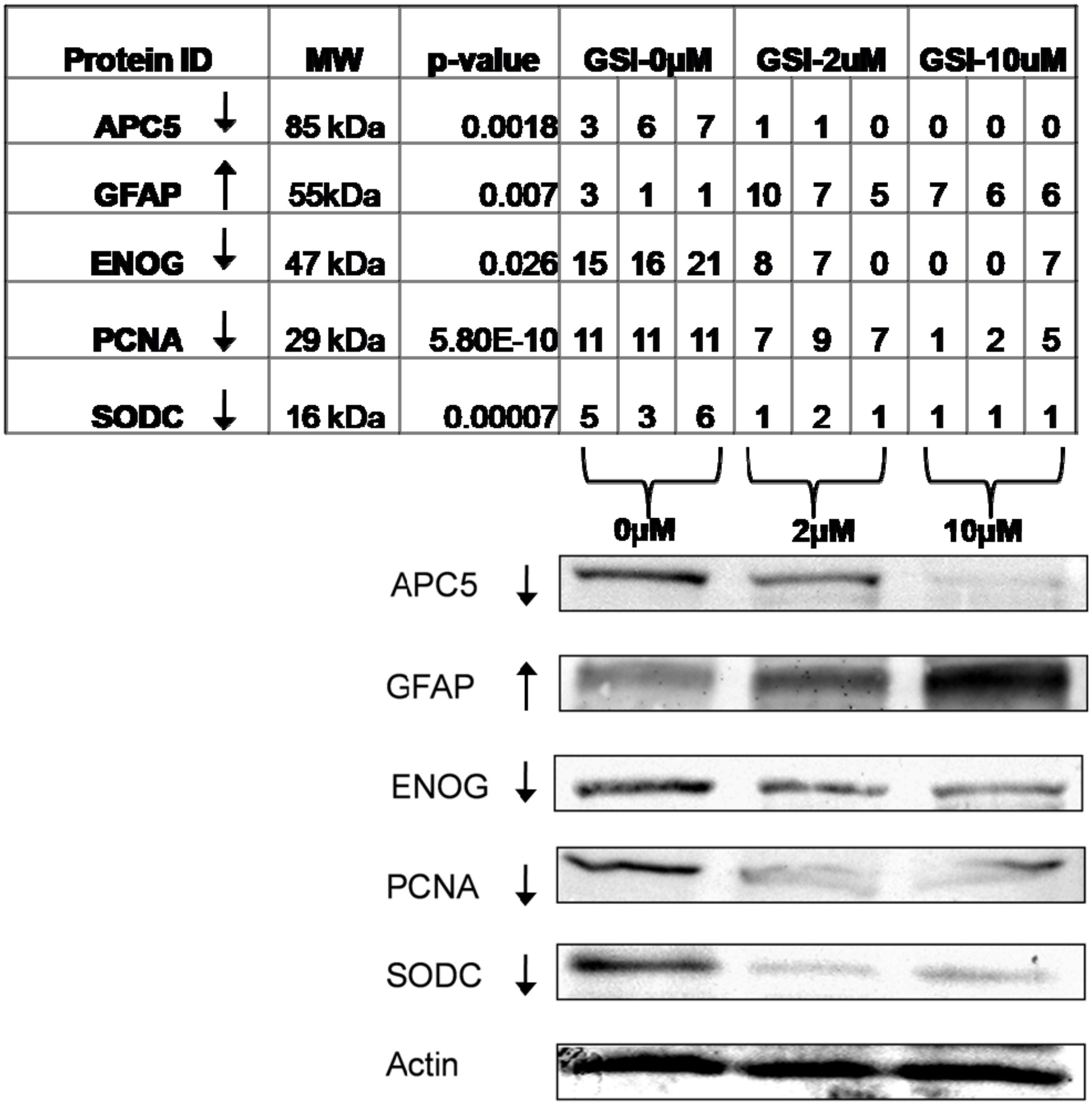


Table 5.1: List of high-priority protein candidates after multilevel filtering. P-value shown here is the averaged from the three pairs student t-tests.

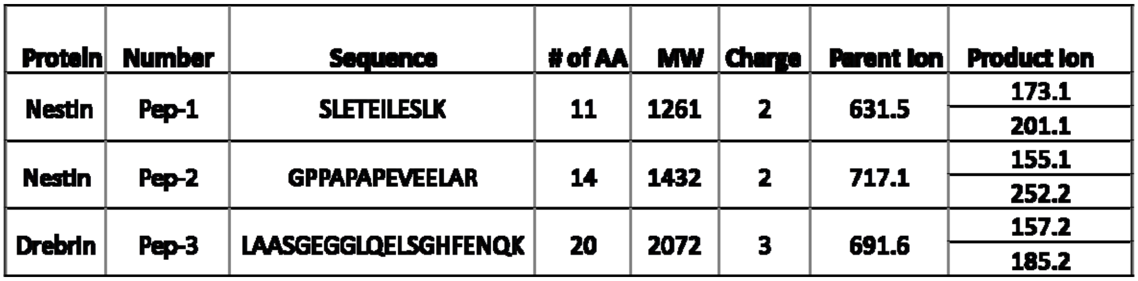
Functions	Accession	Protein Name	Gene Name	P-value*
Proliferation	PCNA_HUMAN	Proliferating cell nuclear antigen	PCNA	0.0016
	NEST_HUMAN	Nestin	NES	0.0067
	PA2G4_HUMAN	Proliferation-associated protein 2G4	PA2G4	0.0080
Differentiation	TBB3_HUMAN	Tubulin beta-3 chain	TUBB3	0.0241
Apoptosis	1433B_HUMAN	14-3-3 protein beta/alpha	YWHAB	0.0028
	VDAC1_HUMAN	Voltage-dependent anion-selective channel protein 1	VDAC1	0.0030
	1433T_HUMAN	14-3-3 protein theta	YWHAQ	0.0053
	1433F_HUMAN	14-3-3 protein eta	YWHAH	0.0116
	PDIA1_HUMAN	Protein disulfide-isomerase	P4HB	0.0093
	TCPB_HUMAN	T-complex protein 1 subunit beta	CCT2	0.0128
	1433E_HUMAN	14-3-3 protein epsilon	YWHAE	0.0139
Tumor Invasion	DREB_HUMAN	Drebrin	DBN1	0.0043
	MYH9_HUMAN	Myosin-9	MYH9	0.0191
Oxidative Response	SODC_HUMAN	Superoxide dismutase [Cu-Zn]	SOD1	0.0238
Glycolysis	KPYM HUMAN	Pyruvate kinase isozymes M1/M2	PKM2	0.0186

Figure 5.3: The upper panel shows the spectral counts detected for each protein across replicate runs within each sample. The lower panel shows the corresponding western blot results. The direction of arrow indicates either up or down-regulated expression.



Protein ID	MW	p-value	GSI-0µM			GSI-2uM			GSI-10uM		
APC5 ↓	85 kDa	0.0018	3	6	7	1	1	0	0	0	0
GFAP ↑	55kDa	0.007	3	1	1	10	7	5	7	6	6
ENOG ↓	47 kDa	0.026	15	16	21	8	7	0	0	0	7
PCNA ↓	29 kDa	5.80E-10	11	11	11	7	9	7	1	2	5
SODC ↓	16 kDa	0.00007	5	3	6	1	2	1	1	1	1

Table 5.2: Summary of the target peptides information and the transitions monitored in MRM.



Protein	Number	Sequence	# of AA	MW	Charge	Parent Ion	Product Ion
Nestin	Pep-1	SLETEILESLK	11	1261	2	631.5	173.1
							201.1
Nestin	Pep-2	GPPAPAPEVEELAR	14	1432	2	717.1	155.1
							252.2
Drebrin	Pep-3	LAASGEGGLQELSGHFENQK	20	2072	3	691.6	157.2
							185.2

Figure 5.4: Extracted chromatogram for all six monitored transitions corresponding to three targeted peptides. Bold box represents two transitions belonging to the same peptide. The last two peptides are eluted with a slight retention time difference of 0.84 second.

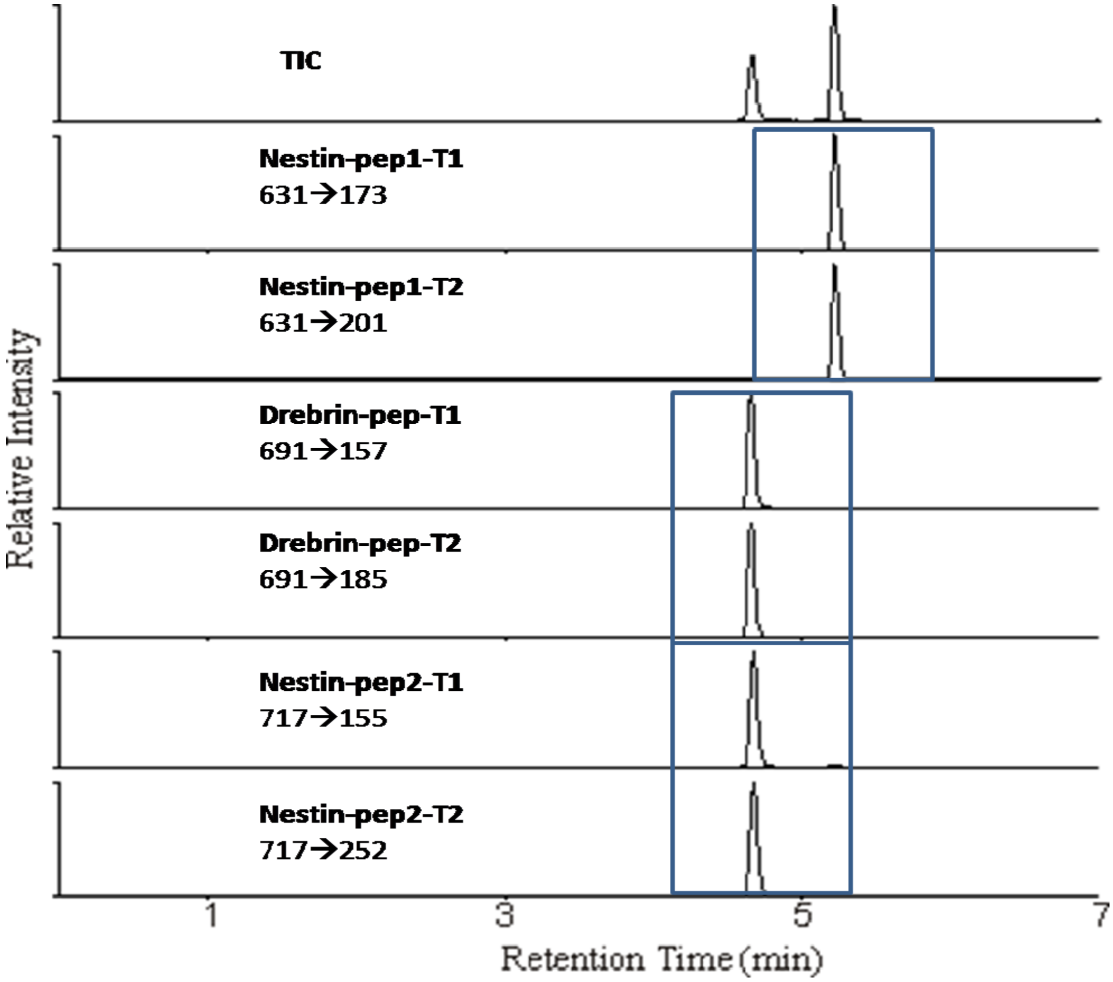
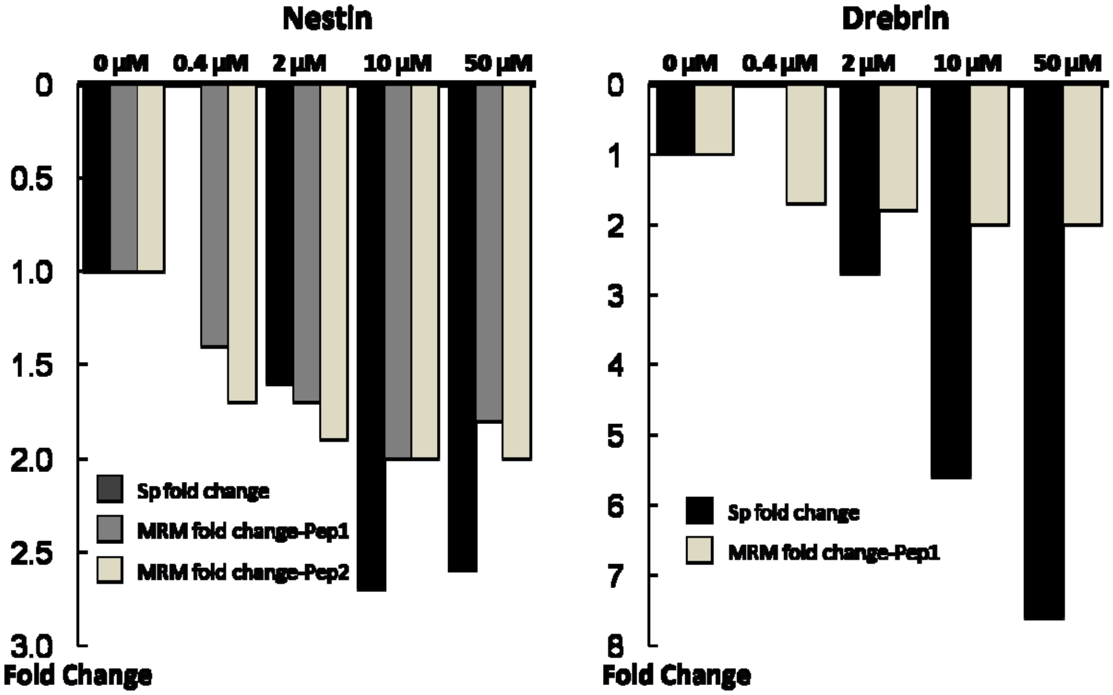
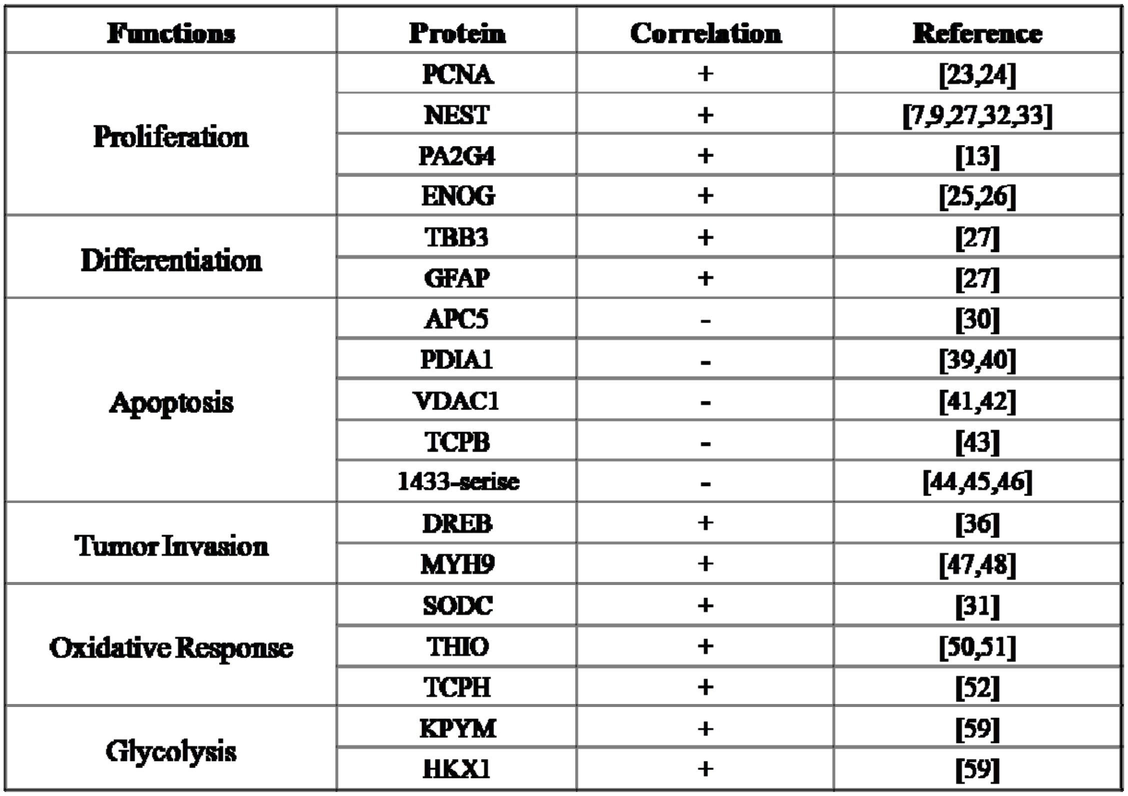


Figure 5.5: Summary of the comparison between the fold change results obtained from label free quantification by LTQ and MRM quantification by QqQ. The y-axis in the upper bar chart represents a ratio of fold changes by dividing the response from each treatment by control. The response of control is normalized to “1”.   The lower panel provides the mean and CV of the spectral counts information for Nestin and Drebrin in each sample. “Spectral Counts” is abbreviated as “Sp”.



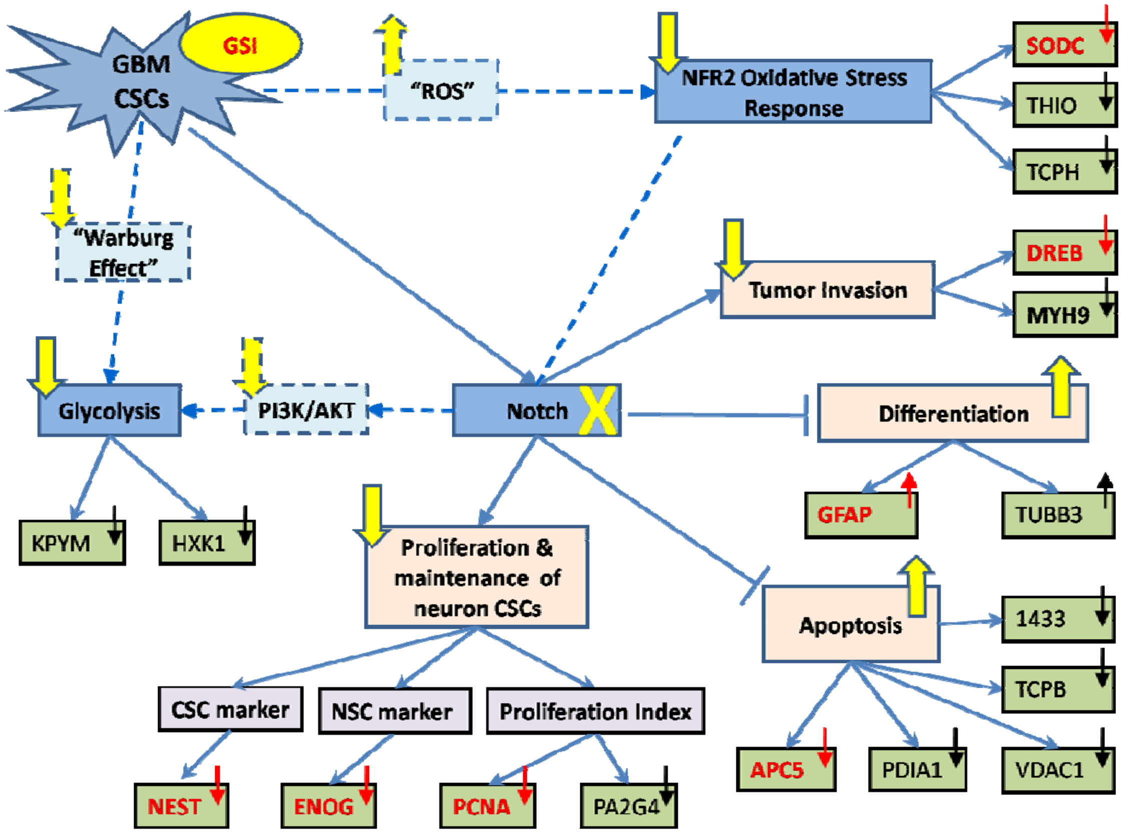
Sp Data	GSI-0µM		GSI-2µM		GSI-10µM		GSI-50µM	
	Mean	CV	Mean	CV	Mean	CV	Mean	CV
Nestin	48	8%	30	15%	18	2%	19	4%
Drebrin	7	4%	3	19%	1	42%	1	3%

Table 5.3: Correlation between the proteins used for constructing the altered signaling network and their targeted functions. "+" indicates positive association and "-" indicates negative association, which are learned from previous literature publications as listed in the "Reference" column.



Functions	Protein	Correlation	Reference
Proliferation	PCNA	+	[23,24]
	NEST	+	[7,9,27,32,33]
	PA2G4	+	[13]
	ENOG	+	[25,26]
Differentiation	TBB3	+	[27]
	GFAP	+	[27]
Apoptosis	APC5	-	[30]
	PDIA1	-	[39,40]
	VDAC1	-	[41,42]
	TCPB	-	[43]
	1433-serise	-	[44,45,46]
Tumor Invasion	DREB	+	[36]
	MYH9	+	[47,48]
Oxidative Response	SODC	+	[31]
	THIO	+	[50,51]
	TCPH	+	[52]
Glycolysis	KPYM	+	[59]
	HKX1	+	[59]

Figure 5.6: Putative Altered Signaling Events occurring upon treatment of GSI. From our current understanding: GBM CSCs signals Notch pathway leading to a constitutive activation. Sequentially, Notch signaling activates the proliferation and maintenance of the undifferentiated CSCs, while suppressing apoptosis and differentiation. Treating GBM CSCs with GSI impairs Notch signaling and therefore reverses the above effects, which is suggested from our results and previous publications. In addition, NFR2-mediated oxidative response and Glycolysis are also suggested to be down-regulated from our results possibly due to their crosstalk with Notch signaling.
Solid line/arrow indicates a conclusion that is drawn based on our results and/or from previous publications with higher confidence. Dashed line/arrow indicates our hypothesis. The direction of an arrow placed in each node represents up/down-regulation of this protein or signaling pathway, while the arrow used to link adjacent nodes represents activation and the blunt end represents inhibition. Blue arrow/line indicates the native state of GBM CSCs while yellow arrow/line indicates the alterations occurring upon GSI treatment. Black arrow indicates the up/down-regulation detected in this study while red arrow indicates the ones that have been further validated by orthogonal approaches in our experiment.



CHAPTER 6

CONCLUSIONS

This dissertation presents the application of quantitative bottom-up MS techniques to analyze biological samples. Differential expression on the protein levels and PTM levels is a critical indication of altered cellular status, which serves as the fundamental motivation for all the thesis projects discussed in the previous chapters. Drawing a correct conclusion regarding a true differential expression requires the synergy from multiple modules: sample preparation, 1D or 2D separation, MS detection and statistical analysis.

Quantitative MS analysis has been firmly established as a powerful tool in the field of differential proteomics. However it also poses several critical challenges. The 2D separation workflow by coupling cIEF with RPLC prior to MS analysis discussed in Chapter2 represents an effort to reduce sample complexity while minimizing the sample usage. The application and optimization of such workflow to analyze clinically important and extremely limited sample (Pancreatic CSCs) is further discussed in Chapter3. A data transformation strategy aiming at resolving the data discontinuity problem embedded in spectral counting-based label free quantitative method is also proposed to gap the bridge in calculating fold changes. Chapter4 switches the focus to a more targeted part of proteomics: the alteration of glycosylation levels in GBM CSCs upon drug treatment. The use of affinity chromatography to enrich the under-represented PTMs is an essential step to reduce sample complexity and purify the fraction of interests. Both lectin microarray and lectin column strategies are employed for this purpose. Student's t-test is the most straightforward statistical analysis to test the significance of differential expression, although it suffers from the limited power to handle nested and hierarchical data structure. Thus, GLMM is explored to model the spectral counting data and the significance is tested from the group effect. There is a certain amount of agreement between t-test and GLMM with one of the important differentially expressed glycoproteins being validated by western blot. Unlike the first 3 chapters which address the technical and/or computational challenges at a particular stage, Chapter5 illustrates a more comprehensive dose-dependent proteomic study that covers a typical biomarker discovery pipeline: global discovery, candidate prioritization and target verification. In addition to the use of western blot as a verification tool, MRM is also employed and the correlation between the label-free and MRM-based quantification results has been demonstrated. Furthermore, biological implications are inferred by data/literature mining. The hypothesis regarding the altered signaling events upon drug treatment in GBM CSCs is proposed, demonstrating the use of quantitative MS techniques to answer essential biological questions.

Overall, quantitative MS techniques are powerful tools to analyze proteome wide differential expressions between paired or multiple biological samples. Future work to ultimately resolve the central challenge caused by high sample complexity may involve further improvements on: 1) separation methods for intact proteins prior to enzymatic digestion; 2) characterization strategies to elucidate detailed structural changes in PTMs; 3) alternative validation tools to perform large scale verifications for global discovery results; 3) computational models to better fit the complicated MS data and increase the inference accuracy of true differential expressions.