Protein Set for Normalization of Spectral Count Data in Quantitative MS Analysis

Wooram Lee

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Biological Sciences

Chair
Iuliana M. Lazar

Carla V. Finkielstein
Yong W. Lee
Jianhua Xing

December 10, 2013
Blacksburg, Virginia

Keywords: proteomics, mass spectrometry, quantitation, normalization

Protein Set for Normalization of Spectral Count Data in Quantitative MS Analysis

Wooram Lee

ABSTRACT

Mass spectrometry has been recognized as a prominent analytical technique for peptide and protein identification and quantitation. With the advent of soft ionization methods, such as electrospray ionization and matrix assisted laser desorption/ionization, mass spectrometry has opened a new era for protein and proteome analysis. Due to its high-throughput and high-resolution character, along with the development of powerful data analysis software tools, mass spectrometry has become the most popular method for quantitative proteomics.

Stable isotope labeling and label-free quantitation methods are widely used in quantitative mass spectrometry experiments. Proteins with stable expression level and key roles in basic cellular functions such as actin, tubulin and glyceraldehyde-3-phosphate dehydrogenase, are frequently utilized as internal controls in biological experiments. However, recent studies have shown that the expression level of such commonly used housekeeping proteins is dependent on cell type, cell cycle or disease status, and that it can change as a result of a biochemical stimulation. Such phenomena can, therefore, substantially compromise the use of these proteins for data validation.

In this work, we propose a novel set of proteins for quantitative mass spectrometry that can be used either for data normalization or validation purposes. The protein set was generated from cell cycle experiments performed with MCF-7, an estrogen receptor positive breast cancer cell line, and MCF-10A, a non-tumorigenic immortalized breast cell line. The protein set was selected from a list of 3700 proteins identified in the different cellular sub-fractions and cell cycle stages of MCF-7/MCF-10A cells, based on the stability of spectral

count data (CV<30 %) generated with an LTQ ion trap mass spectrometer. A total of 34 proteins qualified as endogenous standards for the nuclear, and 75 for the cytoplasmic cell fractions, respectively. The validation of these proteins was performed with a complementary, Her2+, SKBR-3 cell line. Based on the outcome of these experiments, it is anticipated that the proposed protein set will find applicability for data normalization/validation in a broader range of mechanistic biological studies that involve the use of cell lines.

Acknowledgments

I would like to thank my advisor, Dr. Iuliana Lazar for her tremendous understanding, support, guidance, and patience throughout my research. I am thankful to my respected committee members Dr. Carla Finkielstein, Dr. Yong Woo Lee and Dr. Jianhua Xing for their time, help, and support throughout my graduate career and in reviewing my thesis.

I would like to thank Milagros Perez, Yang Xu, Jingren Deng, Fumio Ikenishi and all of my lab mates in the Lazar lab for all of their support and laughter.

Finally, I would like to thank my family and friends for their love and prayers. This work is dedicated to them.

Table of Contents

List of Figures

Chapter 2

Chapter 4

List of Tables

Chapter 2

Chapter 4

Chapter 1. Research Objectives

A number of studies have shown recently that the expression level of commonly used housekeeping proteins is dependent on various factors, such as cell type, cell cycle, disease status and external biochemical stimulation. Therefore, in quantitative biological comparisons, under certain experimental conditions, the use of these proteins as a control or for data validation can substantially alter the interpretation of results and lead to erroneous conclusions. To improve the accuracy of quantitative biological experiments through mass spectrometry detection, and to increase the reliability of normalization and validation by spectral counting, in this work we propose to accomplish the following objectives:

1. Develop a strategy that will enable the identification of endogenous cell line proteins that maintain stable expression level under experimental conditions that induce a major biological perturbation. Two cell lines, one MCF-7 (ER+) breast cancer, and one MCF-10A (non-tumorigenic), will be cultured in two different cell cycle stages (G1 and S) and separated into nuclear and cytoplasmic fractions.

2. Propose a set of proteins that can be used for the normalization/validation of biological data generated by mass spectrometry analysis. Proteins that display minimal variability in their spectral count data across all experimental conditions will be selected and evaluated for suitability for normalization/validation.

3. Validate the proposed protein set. A complementary cell line, SKBR3 (Her2+), will be used to assess the stability in expression level of the proposed protein set.

Chapter 2. Introduction

2.1. Mass spectrometry

2.1.1. Overview of mass spectrometry

Mass spectrometry (MS) is a technique that is used for determining the molecular mass of an analyte by measuring its m/z (mass-to-charge) ratio. Additional applications involve elemental composition determination and elucidating chemical structures. Due to the analyzing power of the instrument, MS has become the most popular method for proteomic studies (**Figure 1**). Single mass spectrometry experiments can enable the identification of up to 15,000 peptides and over 4,000 proteins [1]. These numbers can vary depending on the type of MS instrument, and the sample concentration, abundance and complexity. The three most important parts of a mass spectrometer are the ion source, the mass analyzer and the detector. As the mass spectrometer can detect only ions in the gas phase, the sample needs to be vaporized and ionized. This process occurs in the ion source, the most commonly used methods for sample ionization in proteomics being electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI). The ions are introduced and further separated in a mass analyzer according to their m/z ratios by electromagnetic fields. After detection by an electron multiplier or multichannel plate, the ion signal is converted into an electrical current, and a computer system processes the ion signals and generates a mass spectrum. Data processing is performed by using a variety of computational and bioinformatics tools.

While the capabilities of MS instrumentation are broad, the availability of facilities that can perform advanced proteomic studies lags far behind. In 2008, the Human Proteome Organization (HUPO) created a working group for testing the reproducibility of LC-MS based technology platforms [2]. HUPO distributed a test sample that contained 20 highly purified recombinant human proteins. Of the 27 laboratories, only 1 laboratory reported correctly all tryptic peptides and 7 laboratories reported all 20 proteins. The raw data generated by the majority of the working groups showed, however, sufficient coverage for all 20 proteins. This outcome demonstrates the need for education and proper training for

the use of complex technologies such as MS. The improvement of database and search engines for MS-based research will continue to enhance the accuracy and fidelity of proteomic studies [2].

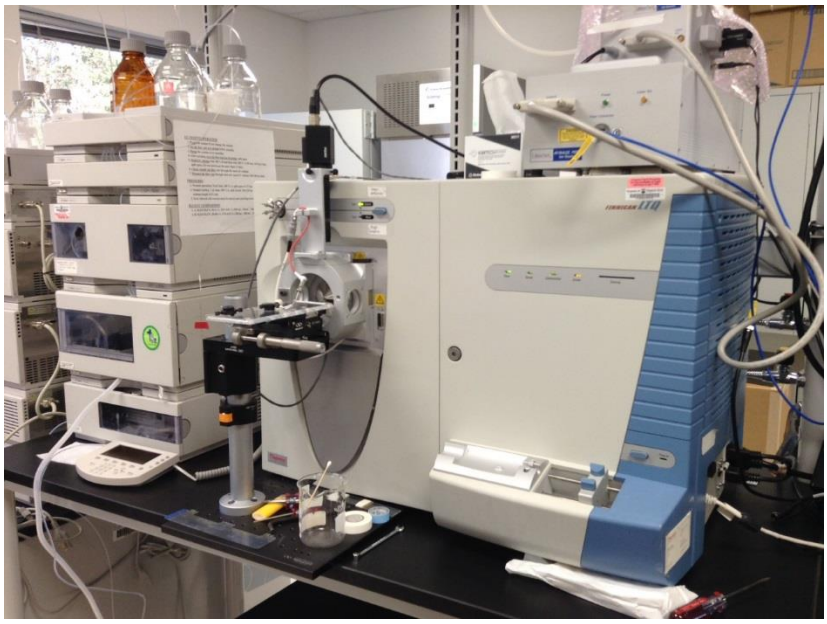


Figure 1. LTQ mass spectrometer and HPLC system.

2.1.2. Ion sources

As a function of molecular structure and solid/liquid or gaseous state of a sample, a variety of ionization methods have been developed for MS detection: electron impact (EI), inductively coupled plasma (ICP), glow discharge, field desorption (FD), plasma desorption (PD), laser desorption (LD), secondary ion mass spectrometry (SIMS), fast atom bombardment (FAB), desorption/ionization on silicon (DIOS), direct analysis in real time (DART), thermospray ionization (TSI) and atmospheric pressure chemical ionization (APCI). Matrix assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are the most widely used methods for the analysis of high molecular weight (MW) biological samples. ESI, in particular, is a powerful tool for the characterization of complex proteomic samples, as it allows the characterization of large MW peptides and proteins without altering the original structure in the ionization step. In an ESI source, the sample peptides are delivered through a capillary and electrosprayed by the generation of

a high electric field between the spraying capillary and the MS. The electrosprayed droplets are diminished in size by solvent evaporation and destabilized by the high density of charge. At the critical point when the electrostatic repulsion forces exceed the surface tension forces, the single droplets explode into smaller, highly charged droplets. This process is repeated until a very fine ion mist is generated and delivered to the ion inlet capillary of the MS system.

2.1.3 Mass analyzers

The mass analyzer is the essential part of a mass spectrometer. It separates the ions according to the m/z value. There are several types of commonly used mass analyzers: sector, time-of-flight (TOF), quadrupole, quadrupole linear ion trap (LIT), quadrupole ion trap (QIT), and Fourier transform ion cyclotron resonance (FT-ICR). Recently developed instruments have a combination of different analyzers to create enhanced MS-MS capabilities. These instruments enable superior mass resolution/accuracy, and the development of novel scanning modes that cannot be performed by a single analyzer. The power of a combined analyzer, together with a flexible data-dependent acquisition routine, has been proven to be of significant importance in proteomics.

Sector. The sector mass analyzer is the classical mass spectrometer. It uses electric and/or magnetic fields that affect the path and velocity of the ionized analyte. Sector instruments curve the trajectories of the ions as they pass through the analyzer, according to their mass-to-charge ratios. Lighter, highly charged ions are affected more by the electromagnetic field than the heavier ions. Scanning over a range of m/z is possible, and specific m/z values can be detected. Sector instruments are bulky and expensive, and are not commonly used in biological analysis. Sector instruments are mainly used for the analysis of small molecules and petroleum samples in environmental and elemental analysis applications via direct-probe and GC-MS.

Time-of-flight. The time-of-flight (TOF) analyzer consists of an ion accelerator, a flight tube and an ion detector. Ions are accelerated by an electrical field in the accelerator region of the mass spectrometer, and released into a field-free flight tube for analysis. After

acceleration, ions acquire a specific amount of kinetic energy. After release in the flight tube, the velocity of the ions will depend on their m/z value. The light ions will fly through the flight tube faster and reach the detector earlier than the heavy ions, if they have the same charge state. The detector will record the arrival time (time-of-flight) for each ion, and a mass spectrum (abundance vs. m/z or flight time) will be generated.

Quadrupole. Quadrupole mass analyzers are made of a set of four parallel rods with a circular or hyperbolic cross section. Ions from the ion source are injected in the space between the rods and subjected to mass analysis in the electrical field created by the DC (direct current) and RF (radio-frequency) voltages that are applied to the quadrupole. A quadrupole mass analyzer can be used as a broad m/z ion transfer device when only the RF field is applied to the rods. However, when both DC and RF fields are applied and scanned, the quadrupole will allow only a specific m/z to pass through the rods. Depending on the amplitude of DC and RF voltages, scanning over a desired m/z range can be accomplished.

Quadrupole ion trap. The quadrupole ion trap (QIT) consists of two end-cap electrodes and one ring electrode. The ring electrode is located between the two end-cap electrodes. Among these three electrodes, DC and RF potentials are applied to trap the ions in a 3D quadrupole field. The ion trap is typically filled with helium in a low pressure system of ~1 mTorr. Collisions with helium gas in the trap promote a contraction of ion trajectories toward the center of the ring electrode. The presence of helium also enables the ejection of ions in dense ion packets during the mass analysis step. By managing the applied potentials to the cell, ions of a particular m/z can be trapped, fragmented or ejected for analysis.

Quadrupole linear ion trap. A quadrupole linear ion trap (LIT) is also an ion trap mass analyzer, similar to a quadrupole ion trap, but it traps ions in a two dimensional quadrupole field, whereas the QIT traps in a three-dimensional quadrupole field. The Thermo Electron LTQ (linear trap quadrupole) mass spectrometer that was used to generate the data in our experiments is an example of a quadrupole linear ion trap. The LIT uses a set of quadrupole rods to trap ions radially. For axial trapping, the LIT has static electrical fields applied to

the end of the quadrupole. The linear trap can store a large number of ions and enables sensitive analysis. It has fast scanning rates and its construction is relatively simple.

Fourier transform ion cyclotron resonance. Fourier transform ion cyclotron resonance (FT-ICR), also known as Fourier Transform Mass Spectrometry (FTMS), excites all the ions present in the cell and detects the composite image current produced by all ions in the ion cyclotron. Ions of a given mass/charge will have a characteristic cyclotron frequency. The image current is converted to individual frequencies generated by single m/z ions by a Fourier transform. The resolution and mass accuracy of FT-ICRMS is significantly higher than that of other types of MS instruments and make them particularly useful for the analysis of posttranslational modifications.

2.2. Proteomics

2.2.1. Proteomics overview

The suffix –ome, as used in molecular biology, refers to a totality of some sort. Therefore, the proteome means the entire set of proteins expressed by the genome, and proteomics is the field of study that concerns itself with the analysis of the proteome. More specifically, proteomics involves the study of the structures, functions and modifications to the proteome. It encompasses not only individual protein identifications but also quantitative measurements of differentially expressed proteins in various biological systems. Due to recent technical advancements in instrumentation and analytical methodologies, the analysis of the proteome characteristic of an organism has improved remarkably. Nevertheless, due to its complexity, proteomic analysis still represents a daunting task. Due to alternative splicing, mRNA editing and post translational modifications, the number of proteins far exceeds the number of genes in an organism. Moreover, low copy number proteins are difficult to detect, especially in the presence of highly abundant proteins. Sensitivity, broad dynamic range, and dynamic composition remain challenges that stimulate researchers in the analysis of the proteome.

2.2.2 Mass spectrometry and tandem mass spectrometry in proteomic analysis

Mass spectrometry was not suitable for proteomics research before the development of ESI and MALDI in the late 1980s. The advent of these two ionization techniques dramatically changed the stand towards proteomics [3] and catalyzed the development of new mass analyzers and complex multi-stage instruments designed to tackle the challenge of protein and proteome analysis. Various types of biological samples can be analyzed by mass spectrometry. The protein samples are first digested into smaller peptides, as the average molecular weight of proteins (30-50 kDa) is too large to be analyzed effectively by most mass spectrometers. However, after proteolytic digestion, the complexity of the sample is increased, and the high complexity often hinders mass analysis. Low abundant signals are suppressed by high abundant signals. Therefore, by adapting a liquid chromatography system before mass analysis, the sample complexity can be reduced, and more information can be extracted from the sample. Additional tandem mass spectrometry analysis can yield detailed results regarding protein post-translational modifications. MS/MS analysis (tandem mass spectrometry) is a multi-step mass spectrometry analysis strategy performed by a series of MS analyzers or a series of MS events. MS/MS enables the generation of amino acid sequence information, a task that single MS cannot perform. Typically, an MS experiment generates data characterizing thousands of proteins and peptides, and to process this large amount of information, bioinformatics tools are used. By using such a workflow, MS instruments can be used for protein identifications, quantitation, detection of PTMs and cellular interactions [4].

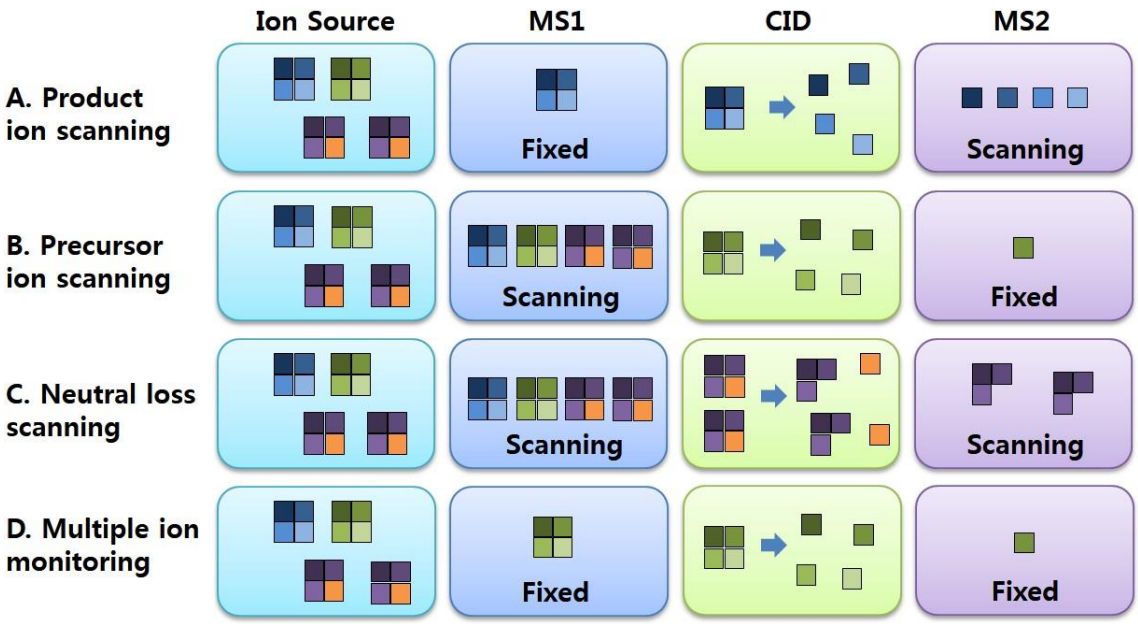


Figure 2. Schematic representation of various types of tandem MS experiments.

Tandem MS can be performed by utilizing a range of scanning methods (**Figure 2**). (A) Product ion scanning is the most common MS/MS experiment in proteomics. Product ion scanning is a procedure that generates a fragment ion spectrum of a precursor ion for the identification of the specific peptide sequence. In this experiment, the first analyzer selects one specific precursor ion, and the selected ion undergoes collision induced dissociation (CID). CID is an ion fragmentation mechanism in which peptide ions collide with neutral molecules, usually helium, and undergo fragmentation. Collision fragments are then analyzed by the second analyzer. (B) Precursor ion scanning is used to detect a subset of peptides in a sample that contain a specific functional group. Precursor ion scanning works the opposite way to product ion scanning. The second analyzer selects a specific daughter ion (fragment ion) and the first analyzer scans for all parent ions (precursor ions). In this experiment, parent ions that can generate specific daughter ions will be detected. (C) Neutral loss scanning scans two analyzers in a synchronized manner, so that the mass difference of ions passing through MS1 and MS2 remains constant. This experiment measures mass differences between parent and daughter ions. Accordingly, neutral loss scans are suitable to detect peptides that contain functional groups of a specific mass such

as a phosphate group. (D) Multiple reaction monitoring (MRM) consists of a series of short experiments in which one precursor ion and one specific fragment characteristic for that precursor are selected by MS1 and MS2, respectively. By MRM, a specific precursor-fragment pair will be detected with better detection limits and improved specificity [3].

2.2.3. Quantitation by mass spectrometry

The identification of proteins and the measurement of protein abundances in biological systems represent tasks of major importance to proteomic studies. Recently, along with the technological developments of MS instrumentation, a variety of quantitative analysis approaches for complex biological samples have emerged. Quantitative proteomics is generally performed by two-dimensional gel electrophoresis, stable isotope labeling and label-free quantitation methods. Two-dimensional gel electrophoresis is the oldest technique, but still commonly used in every-day practice. Label-free quantitation is based on ion intensity or area measurements, or spectral counts, and has shown promising results in proteomics. Due to better accuracy, however, a broad range of quantitative experiments rely on stable isotope labeling, the most commonly used techniques being SILAC (stable isotope labeling with amino acids in cell culture), ICAT (isotope-coded affinity tags) and iTRAQ (isobaric tags for relative and absolute quantitation). Alternative methods such as MCAT (mass-coded abundance tagging) or NIT (N-terminal isotope encoded tagging), however, have been used by some MS communities to better address particular biological applications.

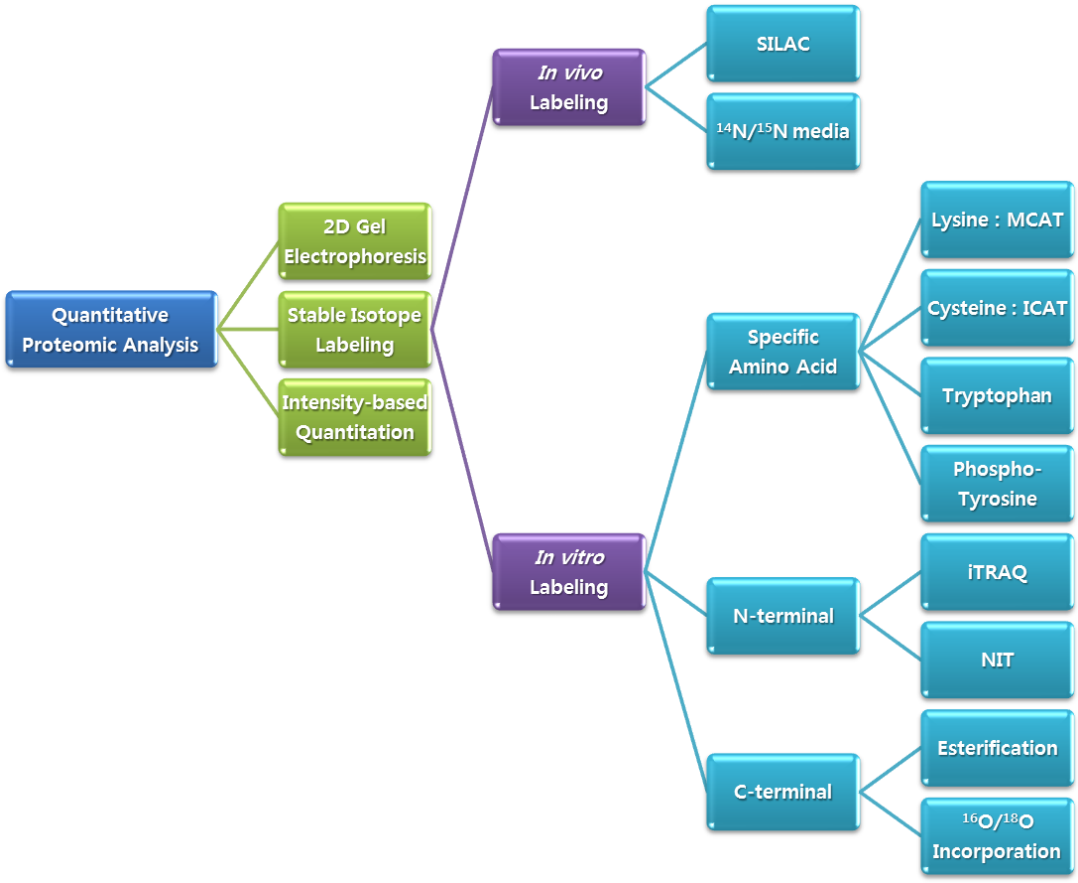


Figure 3. Quantitative proteomic analysis. SILAC: stable isotope labeling with amino acids in cell culture; MCAT: mass-coded abundance tagging; ICAT: isotope-coded affinity tags; iTRAQ: isobaric tags for relative and absolute quantitation; NIT: N-terminal isotope encoded tagging [5].

SILAC. The most popular in vivo stable isotope labeling technique is stable isotope labeling with amino acids in cell culture (SILAC) [6]. The method relies on labeling the proteins in the growing cells by media that contain isotopically labeled amino acids (e.g., Lys). As labeling occurs during cell proliferation, the experimental variability that is introduced usually by in vitro labeling techniques is eliminated. Therefore, quantitation accuracy is high. For quantitation, proteins extracted from cells cultured in light medium are compared to proteins extracted from cells cultured in heavy medium. SILAC is not amenable, however, for analyzing tissues or body fluids, which annot grow in isotopically labeled media.

ICAT. In vitro stable isotope labeling was introduced by Gygi et al. in 1999 [7]. The method was termed 'isotope-coded affinity tags (ICAT)' and is a cysteine amino acid specific isotope labeling technique. The ICAT reagent consists of a thiol-specific reactive group, an isotope labeled linker and a biotin affinity group. The linker contains heavy or light isotopes of hydrogen to label two different samples. The cysteines in proteins in two different samples are labeled with either the light or the heavy ICAT reagents. Differential expression is evaluated based on the areas or intensities of ions corresponding to the labeled peptides. Unlike in vivo labeling, ICAT is amenable for labeling tissues or body fluids, and also cultured cells.

iTRAQ. As only ~96 % of proteins and ~27 % of tryptic peptides in the human proteome contain cysteine residues, the ICAT technology is unable to cover the entire human proteome. In order to acquire 100 % coverage, other N-terminal or C-terminal peptide labeling techniques must be used. Recently, novel Lys/N-terminal isotope labeling technologies, such as iTRAQ (isobaric tags for relative and absolute quantitation), have been developed for peptide quantitation [8]. The iTRAQ reagents consists of a reporter group (mass ranging from 114 to 117 Da), a balance group (mass ranging from 31 to 28 Da) and an amine-specific peptide-reactive group (NHS). The reagents can be used for 4-plex or 8-plex labeling experiments. The mass of the sum of the reporter and balance groups is 145 Da (114+31, 115+30, 115+29, 117+28) for all four reagents (for the case of a 4-plex reagent set). Protein digests from different samples are labeled with iTRAQ reagents with different tags and are mixed. During collision induced dissociation in the mass analyzer, the reporter group falls apart from the peptide, displaying a distinct mass of 114 to 117 Da. The intensity of the fragment reporter groups is used for peptide quantitation in multiple samples. The remaining peptide also fragments through CID to provide amino acid sequence information for peptide identifications.

Label-free techniques. Label-free techniques are, at present, in the limelight as a promising alternatives to stable-isotope labeling techniques. As a result, the number of publications that describe label-free approaches for differential protein expression analysis has increased substantially in the last decade. The major advantage of label-free

quantitation include: (i) the sample alterations are minimized, (ii) the workflow is straightforward, (iii) the expense of sample preparation is reduced, and (iv) there are no limitations in number of samples that can be directly compared in a study. Even though the accuracy of the quantitative data is generally inferior to that of the data generated by isotope-labeling methods, well-designed experiments can lead to information-rich and reliable results on a broad biological scale [9]. Label-free quantitation is performed by various methods that include the absolute mass tag (AMT) approach, total ion chromatography (TIC), peptide ion intensity, fragment ion intensity, absolute protein expression (APEX), area under the curve (AUC), and spectral counting [10-13].

2.2.4. Data normalization in quantitative MS analysis

To decrease the impact of biological, experimental and technical variability that can reach multi-fold values in un-properly designed biological experiments, data normalization is absolutely imperative in differential expression proteomic studies [14]. The efficacy of the data normalization can be assessed by using a simple and widely used indicator, such as the coefficient of variation (CV). For example, in the analysis of a whole proteome digest of Streptococcus pyogenes, it was shown that by using four stable ribosomal proteins for data normalization [9], the normalization process was able to reduce the CV values by up to 75 %, to levels of ~20 % [9]. Optimizing the parameters of search algorithms such as Mascot, MS-Fit, Profound and SEQUEST could further improve the results [15].

Both internal and external standards can be used for the normalization. External standards are proteins that are spiked into a sample to assess experimental and technical errors. Internal standards are housekeeping gene products, mRNAs and proteins that can be used as an endogenous control in differential expression studies (**Table 1**). Housekeeping genes are constitutive genes that are expressed in all cells at rather constant level under both normal and altered conditions, and their role is to maintain fundamental cellular functions. Some of the most commonly used examples include: actin, tubulin and GAPDH (glyceraldehyde-3-phosphate dehydrogenase). Their utility as universal standards has been questioned, however, and careful selection based on the specifics of a given biological

experiment, and the nature of the sample or tissue to be analyzed, was suggested instead. Housekeeping mRNAs are typically used for normalization in RT-PCR, RNase protection assay and qPCR experiments, while protein products are used for studies that involve 2D gel electrophoresis, western blot and mass spectrometry experiments. Their utility was tested in a broad range of tissues, organs or cell lines, and some of the gene products were also found to be useful subcellular and organelle markers of the nucleus, peroxisome, cytoplasm, ribosome, ER and mitochondria.

Table 1. Housekeeping genes and gene products used for data normalization in quantitative differential expression studies.

Protein Name	Detection Level	Tissue, Disease	Cellular Location	Experiment Type	Ref.
18S rRNA*	mRNA	Brain (M)		RT-PCR	[16]
28S rRNA*	mRNA	Spleen (M)		RNase protection assay	[16]
ACTB (Beta-actin)*	Protein	Liver, hepatocellular carcinoma (H)		2-DE, Western Blot, qPCR	[17]
ACTB (Beta-actin)*	Protein	Omental fat cell (H)		2-DE, Western Blot, MS (Ultraflex MALDI-TOF)	[18]
ACTB (Beta-actin)*	mRNA	Spinal cord, Brain, Skeletal muscle (M)		RT-PCR	[19]
Beta-Tubulin*	Protein	Brain, Skeletal muscle (M)		Western Blot	[19]
ENOA (Enolase I)*	Protein	Omental fat cell (H)		2-DE, Western Blot, MS (Ultraflex MALDI-TOF)	[18]
GAPDH*	mRNA	Brain, Skeletal muscle (M)		RT-PCR	[19]
GAPDH*	Protein	Spinal cord, Brain (M)		Western Blot	[19]
GAPDH*	mRNA	Brain (M)		RT-PCR	[16]
HSP60 (Heat Shock Protein 60)*	Protein	Liver, hepatocellular carcinoma (H)		2-DE, Western Blot, qPCR	[17]
L32 (60S ribosomal protein L32)*	mRNA	Brain (M)		RT-PCR	[16]
PARK7 (Parkinson disease protein 7)*	Protein	Omental fat cell (H)		2-DE, Western Blot, MS (Ultraflex MALDI-TOF)	[18]
PDI (Protein Disulphide Isomerase)*	mRNA, Protein	Liver, hepatocellular carcinoma(H)		2-DE, Western Blot, qPCR	[17]
ACTB (Beta-actin)	Protein	Breast cancer (H)		Mass Spec (LTQ-Orbitrap)	[20]
Apoa1 (Apolipoprotein A-I)	Protein	Breast cancer (M)		Mass spec (QTRAP 5500)	[21]
Apoa4 (Apolipoprotein A-IV)	Protein	Breast cancer (M)		Mass spec (QTRAP 5500)	[21]
CALR (Calreticulin)	Protein	Macrophage (M)	ER	Mass Spec (TSQ Vantage)	[22]
Cpn2 (Carboxypeptidase N, polypeptide 2)	Protein	Breast cancer (M)		Mass spec (QTRAP 5500)	[21]

DDX3 (DEAD box proteins 3)	Protein	Lung cancer (H)	Nucleus	Western Blot	[23]
ES1 (ES1 protein homolog)	Protein	Breast cancer (M)	Mitochondria	Mass spec (QTRAP 5500)	[21]
GAPDH	Protein	Macrophage (M)		Mass Spec (TSQ Vantage)	[22]
Gsn (Isoform 1 of Gelsolin)	Protein	Breast cancer (M)		Mass spec (QTRAP 5500)	[21]
HDAC1 (Histone deacetylases 1)	Protein	Lung cancer (H)	Nucleus	Western Blot	[23]
HDAC2 (Histone deacetylases 2)	Protein	Kidney, lung, liver (M)	Nucleus	Western Blot	[23]
HSP90 (Heat Shock Protein 90)	Protein	Lung cancer (H)	Cytoplasm	Western Blot	[23]
HSPD1 (HSP60, mitochondrial)	Protein	Macrophage (M)	Mitochondria	Mass Spec (TSQ Vantage)	[22]
Itih1 (Inter-alpha-trypsin inhibitor heavy chain 1)	Protein	Breast cancer (M)		Mass spec (QTRAP 5500)	[21]
LDHA (Lactate dehydrogenase A)	Protein	Macrophage (M)	Peroxisome, Cytoplasm	Mass Spec (TSQ Vantage)	[22]
MCM2 (Minichromosome maintenance complex component 2)	Protein	Lung cancer (H)	Nucleus	Western Blot	[23]
MEK1 (Mitogen-activated protein kinase kinase 1)	Protein	Lung cancer (H)	Cytoplasm	Western Blot	[23]
MSH2 (MutS homolog 2)	Protein	Lung cancer (H)	Nucleus	Western Blot	[23]
NCL (Nucleolin)	Protein	Macrophage (M)	Nucleus	Mass Spec (TSQ Vantage)	[22]
p53	Protein	Lung cancer (H)	Nucleus	Western Blot	[23]
Pzp (Pregnancy zone protein)	Protein	Breast cancer (M)		Mass spec (QTRAP 5500)	[21]
Ran (RAs-related Nuclear protein)	Protein	Lung cancer (H)	Nucleus	Western Blot	[23]
RPS8 (40S Ribosomal protein S8)	Protein	Macrophage (M)	Ribosome		[23]
SP1 (Specificity Protein 1)	Protein	Heart, lung, liver (M)	Nucleus	Western Blot	[23]
TOPO II beta (Topoisomerases 2 beta)	Protein	Lung cancer (H)	Nucleus	Western Blot	[23]

For example, 28S rRNA and 18S rRNA were recommended as internal mRNA standards for studies of rat brain by RT-PCR, and of mouse spleen and human peripheral blood mononuclear cells by RNase protection assays [16]. In the study of an amyotrophic lateral sclerosis mouse model, beta-actin and GAPDH mRNA were found as suitable housekeeping genes for RT-PCR studies of the skeletal muscle and brain, whereas the beta-actin and GAPDH proteins were found suitable for spinal cord and brain studies by western blotting [19]. The beta-tubulin protein was suggested for brain studies, as well. Other experiments validated beta-actin and heat shock protein 60 at both protein and mRNA level for the study of human hepatic tissues and hepatocellular carcinoma by western blot, immunohistochemistry and real-time quantitative PCR [17]. For the case of an adipose tissue analysis of omental and subcutaneous fat depots, PARK7 (Parkinson disease protein 7), ENOA (Enolase I) and beta-actin were proposed as proper reference standards by western blot [18].

However, there is increasing body of evidences that suggests that commonly used housekeeping proteins are not actually universal standards, but rather cell line specific [24, 25]. As the "one-size-fits-all" internal marker does not exist so far, there is a need for identifying larger sets of endogenous proteins that could be used as a whole, with greater confidence, in the normalization of differential expression biological data, or sub-sets, that are cell-type or disease specific standards, for quantitative MS proteomics research [26].

2.3. References

1. de Godoy, L. M., J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther, and M. Mann. *Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast.* Nature, 2008. 455(7217): p. 1251-4.

2. Bell, A. W., E. W. Deutsch, C. E. Au, R. E. Kearney, R. Beavis, S. Sechi, T. Nilsson, J. J. Bergeron, and Hupo Test Sample Working Group. *A HUPO test sample study reveals*

common problems in mass spectrometry-based proteomics. Nat Methods, 2009. 6(6): p. 423-30.

3. Domon, B. and R. Aebersold. *Mass spectrometry and protein analysis.* Science, 2006. 312(5771): p. 212-7.

4. Walther, T.C. and M. Mann. *Mass spectrometry-based proteomics in cell biology.* J Cell Biol, 2010. 190(4): p. 491-500.

5. Yan, W. and S.S. Chen. *Mass spectrometry-based quantitative proteomic profiling.* Brief Funct Genomic Proteomic, 2005. 4(1): p. 27-38.

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

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

8. Ross, P. L., Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, et al. *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Mol Cell Proteomics, 2004. 3(12): p. 1154-69.

9. Teleman, J., C. Karlsson, S. Waldemarson, K. Hansson, P. James, J. Malmstrom, and F. Levander. *Automated selected reaction monitoring software for accurate label-free protein quantification.* J Proteome Res, 2012. 11(7): p. 3766-73.

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

11. Old, W. M., K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing, and N. G. Ahn. *Comparison of label-free methods for quantifying human proteins by shotgun proteomics.* Mol Cell Proteomics, 2005. 4(10): p. 1487-502.

12. Wang, G., W. W. Wu, W. Zeng, C. L. Chou, and R. F. Shen. *Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes.* J Proteome Res, 2006. 5(5): p. 1214-23.

13. Podwojski, K., M. Eisenacher, M. Kohl, M. Turewicz, H. E. Meyer, J. Rahnenfuhrer, and C. Stephan. *Peek a peak: a glance at statistics for quantitative label-free proteomics.* Expert Rev Proteomics, 2010. 7(2): p. 249-61.

14. Cappadona, S., P. R. Baker, P. R. Cutillas, A. J. Heck, and B. van Breukelen. *Current challenges in software solutions for mass spectrometry-based quantitative proteomics.* Amino Acids, 2012. 43(3): p. 1087-108.

15. Chamrad, D. C., G. Korting, K. Stuhler, H. E. Meyer, J. Klose, and M. Bluggel. *Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data.* Proteomics, 2004. 4(3): p. 619-28.

16. Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., Heinen, E. *Housekeeping genes as internal standards: use and limits.* J Biotechnol, 1999. **75**(2-3): p. 291-5.

17. Sun, S., Yi, X., Poon, R. T., Yeung, C., Day, P. J., Luk, J. M. *A protein-based set of reference markers for liver tissues and hepatocellular carcinoma.* BMC Cancer, 2009. **9**: p. 309.

18. Perez-Perez, R., Lopez, J. A., Garcia-Santos, E., Camafeita, E., Gomez-Serrano, M., Ortega-Delgado, F. J., Ricart, W. *Uncovering suitable reference proteins for expression studies in human adipose tissue with relevance to obesity.* PLoS One, 2012. **7**(1): p. e30326.

19. Calvo, A. C., Moreno-Igoa, M., Manzano, R., Ordovas, L., Yague, G., Olivan, S., Munoz, M. J., Zaragoza, P., Osta, R. *Determination of protein and RNA expression levels of common housekeeping genes in a mouse model of neurodegeneration.* Proteomics, 2008. **8**(20): p. 4338-43.

20. Bateman, N. W., Sun, M., Hood, B. L., Flint, M. S., Conrads, T. P. *Defining central themes in breast cancer biology by differential proteomics: conserved regulation of cell spreading and focal adhesion kinase.* J Proteome Res, 2010. **9**(10): p. 5311-24.

21. Whiteaker, J. R., Lin, C., Kennedy, J., Hou, L., Trute, M., Sokal, I., Yan, P., Schoenherr, R. M., Zhao, L., Voytovich, U. J., Kelly-Spratt, K. S., Krasnoselsky, A., Gafken, P. R., Hogan, J. M., Jones, L. A., Wang, P., Amon, L., Chodosh, L. A., Nelson, P. S., McIntosh, M. W., Kemp, C. J., Paulovich, A. G. *A targeted proteomics-based pipeline for verification of biomarkers in plasma.* Nat Biotechnol, 2011. **29**(7): p. 625-34.

22. Kinter, C. S., Lundie, J. M., Patel, H., Rindler, P. M., Szweda, L. I., Kinter, M. *A quantitative proteomic profile of the Nrf2-mediated antioxidant response of macrophages to oxidized LDL determined by multiplexed selected reaction monitoring.* PLoS One, 2012. **7**(11): p. e50016.

23. Bomgarden, R.D., M. McGirk, and R. Farooquoi. *Nuclear and cytoplasmic protein fractionation from tissue*. Available from: http://www.piercenet.com/product/ne-per-nuclear-protein-extraction-kit.

24. Ferguson, R. E., H. P. Carroll, A. Harris, E. R. Maher, P. J. Selby, and R. E. Banks. *Housekeeping proteins: a preliminary study illustrating some limitations as useful references in protein expression studies.* Proteomics, 2005. 5(2): p. 566-71.

25. Sheng, W.Y. and T.C. Wang. *Proteomic analysis of the differential protein expression reveals nuclear GAPDH in activated T lymphocytes.* PLoS One, 2009. 4(7): p. e6322.

26. Xie, F., T. Liu, W. J. Qian, V. A. Petyuk, and R. D. Smith. *Liquid chromatography-mass spectrometry-based quantitative proteomics.* J Biol Chem, 2011. 286(29): p. 25443-9.

Chapter 3. Materials and Methods

Materials. MCF-7 breast cancer and MCF-10A non-tumorigenic breast epithelial cells, Eagle's minimum essential medium (EMEM), 0.25 % trypsin/0.53 mM EDTA solution, phosphate-buffered saline (PBS) and cell culture grade water were purchased from the American Tissue Culture Collection (Manassas, VA). Fetal bovine serum (FBS) was obtained from Gemini Bio-products (West Sacramento, CA) and sequencing-grade modified trypsin was acquired from Promega Corporation (Madison, WI). Bovine pancreas insulin solution, 17-β estradiol, L-glutamine, Cell Lytic™ NuCLEAR™ extraction kit, phosphatase inhibitors (Na_3VO_4 and NaF), dithiothreitol (DTT), urea, acetic acid, trifluoroacetic acid, ammonium bicarbonate and bovine protein standards (hemoglobin α/β, carbonic anhydrase, α-lactalbumin, fetuin, α-casein, β-casein and cytochrome c) were purchased from Sigma (St. Louis, MO). Phenol-red free Dulbecco's modified Eagle's medium (DMEM) was from Life Technologies (Carlsbad, CA) and charcoal/dextran treated FBS from Hyclone (Logan, UT). SPEC-PTC18 and SPEC-PTSCX solid-phase extraction pipette tips were purchased from Varian Inc. (Lake Forest, CA), and HPLC-grade methanol and acetonitrile from Fisher Scientific (Fair Lawn, NJ). Water was either deionized with a MilliQ Ultrapure water system (Millipore, Bedford, MA), or distilled in house.

Cell culture. MCF-7 and MCF-10 cells were cultured in an incubator at 37℃ (5 % CO_2). The culture medium was EMEM supplemented with FBS (10 %) and bovine insulin (10 μg/ml) for MCF-7, and DMEM:nutrient mixture F-12 (1:1) supplemented with 5 % horse serum, 20 ng/mL hEGF, 0.5 μg/mL hydrocortisone, 0.1 μg/mL cholera toxin and 10 μg/mL bovine insulin, for MCF-10. After several passages, the cells were arrested in the G1 stage by serum deprivation for 48 h in DMEM with 4 mM L-glutamine (MCF-7), or in DMEM/F12 (MCF-10). After arrest, the cells were released into the S phase by a 24 h treatment with DMEM with 1 nM E2, 10 % charcoal/dextran-treated FBS, 4 mM L-glutamine, and 10 μg/mL bovine insulin (MCF-7), or MCF-10 culture medium with 10 % horse serum (MCF-10). Cells were detached from the flask by treatment with trypsin-EDTA solution (0.25 % trypsin, 0.53 mM EDTA), rinsed with PBS (pH 7.4), harvested

and stored in a - 80℃ freezer. The entire process was repeated for three biological replicates. Each replicate sample was analyzed by fluorescent activated cell sorting (FACS) conducted on a Beckman Coulter EPICS XL-MCL analyzer (Brea, CA, USA).

Cell processing. Before MS analysis, the cells were thawed from -80℃, lysed, and separated into nuclear and cytoplasmic fractions by using the Cell Lytic™ NuCLEAR™ extraction kit. First, the cells were incubated for 15 min in hypotonic buffer (10 mM HEPES, pH 7.9, with 1.5 mM $MgCl_2$, 10 mM KCl) supplemented with DTT to a final concentration of 0.01 M, protease inhibitor cocktail and phosphatase inhibitors. IGEPAL CA-630 was added after incubation to a final concentration of 0.6 % (v/v), and the sample was vigorously vortexed for 10 seconds. The sample was centrifuged for 30 seconds at 10,500 x g, and the supernatant, which was the cytoplasmic fraction, was collected and stored on ice. The pellet that contained the nuclear fraction was reconstituted in extraction buffer (20 mM HEPES, pH 7.9, with 1.5 mM $MgCl_2$, 0.42 M NaCl, 0.2 mM EDTA and 25 % glycerol (v/v)), supplemented with DTT to a final concentration of 0.01 M, protease inhibitor cocktail and phosphatase inhibitors. The mixture was vortexed at medium speed for 45 min while avoiding foam formation. The sample was centrifuged for 5 min at 20,500 x g, and the supernatant that contained the nuclear proteins was collected and stored on ice. After nuclear/cytoplasmic separation, protein concentrations were measured by the Bradford assay (SmartSpec Plus spectrophotometer, Bio-Rad, Hercules, CA). The concentration of the protein extracts was adjusted to 5 mg/ml, the samples were denatured, reduced with 8 M urea and 4.5 mM DTT (1 h, 60℃). After a tenfold dilution with 50 mM NH4HCO3, the extract was spiked with an eight standard bovine protein mixture (5 µM each), digested with trypsin for 24 h in 37℃, quenched with glacial CH_3COOH, and cleaned up with C18/SCX cartridges. After clean-up, each sample was reconstituted with CH_3CN/H_2O/TFA solution (5:95:0.1) to a final concentration of approximately 2 µg/µL in proteins and 0.2 µM bovine standards.

LC-MS analysis. From each sample, five technical replicates were analyzed by an Agilent 1100 micro HPLC system (Agilent Technologies, Palo Alto, CA) coupled with a linear trap quadrupole (LTQ) mass spectrometer (Thermo Electron Corporation, San Jose, CA), using

an on-column/no split injection set up. The reverse-phase liquid chromatography (RPLC) separation columns were prepared in-house by packing 100 μm i.d. x 12 cm fused silica capillaries with 5 μm Zorbax SB-C18 particles (Agilent Technologies), and operated at ~160-180 nL/min flow rate. About 1 cm long capillary measuring 20 μm i.d. x 90 μm o.d. was inserted into the RPLC separation column to generate a nanospray emitter. Mobile phases A and B were composed of H_2O:CH_3CN:TFA in 95:5:0.01 and 20:80:0.01 v/v ratios, respectively. The separation gradient (from 0 % to 100 % B) was 3 h long. Each MS scan was followed by zoom scan and MS/MS scans on the five most intense peaks. The parameters used for analysis were: ±5 m/z zoom scan width, dynamic exclusion at repeat count of 1, repeat duration of 30 s, exclusion list size of 200, exclusion duration of 60 s, and ±1.5 m/z exclusion mass width. Tandem MS parameters were: isolation width 3 m/z, normalized collision energy 35 %, activation Q 0.25, and activation time 30 ms.

Chapter 4. Results and Discussion

4.1 Requirements for ideal proteins suitable for normalization purposes.

Protein quantitation by MS analysis is typically performed at the peptide level. The life of a protein is initiated by an extra- or intra-cellular signal that induces DNA transcription and translation. Proteins are then synthesized in the ribosomes/ER and delivered to specific locations in the cell such as the nucleus, mitochondria, Golgi apparatus or cell membrane.

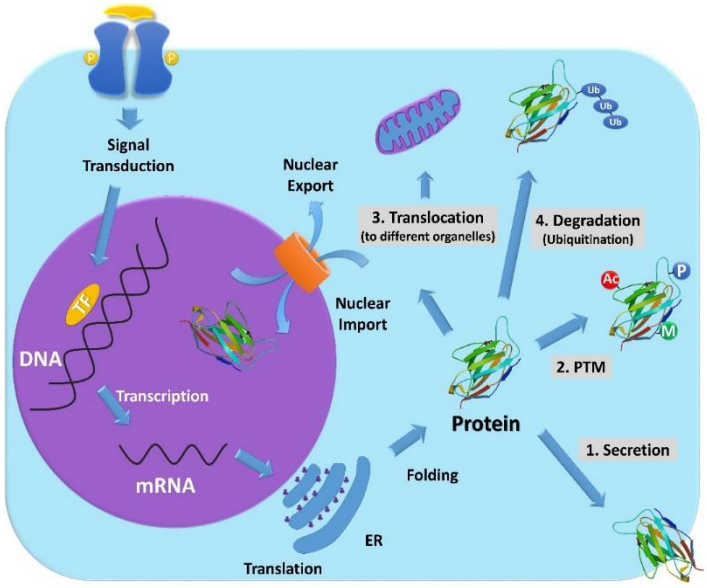


Figure 1. Life of a protein.

Such proteins can be subjected to further processes that result in sub-cellular relocation, ubiquitination and degradation, modifications by PTMs to fulfill certain biological functions, or secretion in the extracellular environment (**Figure 1**). To be used as internal standards for data normalization, ideally, cellular proteins should satisfy a number of requirements.

(a) The expression level of these proteins should remain constant irrespective of the biological experiment that is performed in the study. As most experiments (gene knockouts, cell transfections, cell stimulations, etc.) are conducted to observe the effect of a perturbation on a particular biological process, the ultimate result will be the up- or down-regulation of certain genes and their associated products. Proteins that are involved in

maintaining routine cellular functions (i.e., housekeeping proteins) are expected, however, to not react to the perturbation and preserve an unchanged expression level (**Figure 2**).

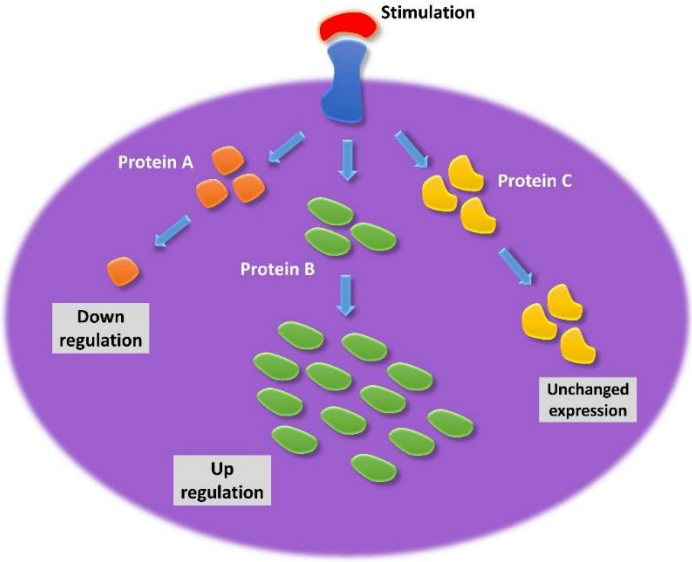


Figure 2. Housekeeping proteins display constant expression level.

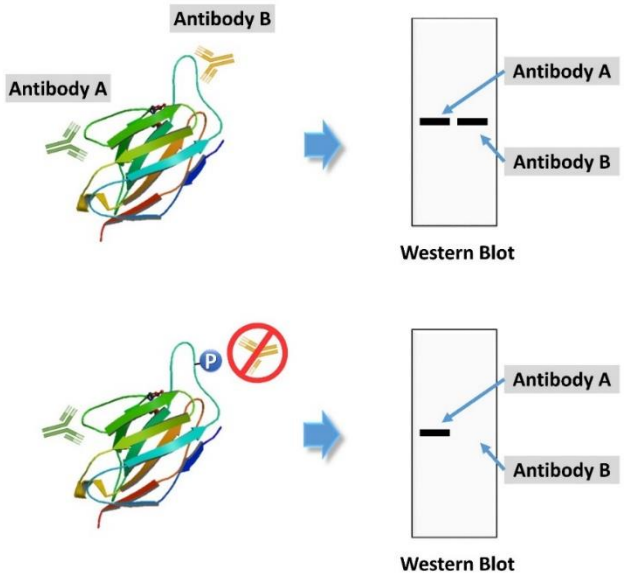
(b) The cellular location of these proteins should be in accordance with their function and in line with data provided by classical studies. The processes that control protein localization and translocation are tightly regulated, as proper protein localization is important for adequate function in a particular physiological context, cell survival and proliferation. Protein mutations, altered expression of cargo and/or transport proteins, deregulation of the protein trafficking machinery, can result, however, in aberrant protein localization [1]. Such miss locations are known to be related to many metabolic, cardiovascular, cancer and neurodegenerative diseases. About 1.5 % of the proteins in glioma, for example, are believed to be miss located as a result of the disease [2]. The majority of proteins, however, including the housekeeping proteins, do not change location, but function within a given spatiotemporal context as part of tissue-specific interaction networks [3].

Table 1. Most frequent protein posttranslational modifications.

PTMs	Related enzymes	Target amino acid
PHOSPHORYLATION	Kinase, Phosphatase	Ser, Thr, Tyr, Arg, Lys, His, Asp, Cys
GLYCOSYLATION	Glycosyltransferase, Glycosidase	Asn, Ser, Thr, Trp, HO-Lys, HO-Pro
ACETYLATION	Acetyltransferase, Deacetylase	Lys, N-terminal
METHYLATION	Methyltransferase, Demethylase	Lys, Arg, His, Glu/Gln, Asp, Cys, N/C-terminal
UBIQUITINATION	E1, E2, E3 enzyme, Deubiquitinating enzyme	Lys

(c) The proteins or peptides that are used for normalization purposes should be free of PTMs. Most proteins carry not just one, but several PTMs, which have the important role of determining protein function, location and fate (**Table 1**). Among the hundreds of known PTMs, the most common ones include phosphorylation, glycosylation, acetylation, methylation and ubiquitination. The covalent attachment and removal of these PTMs to target amino acids occurs through reversible reactions catalyzed by specific enzymes. PTM status can change, however, as a result of the biological perturbations that are induced during a study. For example, GAPDH, a key enzyme involved in glycolysis that is often used as an endogenous control, is primarily located in the cytoplasm. It can translocate, however, to the nucleus following S-nitrosylation on Cys-152 and interaction with SIAH2 [4], and it was also found localized at the cell membrane, polysomes, ER and Golgi [4]. Its cellular location is also dependent on whether the cells are cycling or non-cycling [5]. Phosphorylation, acetylation and ubiquitination affect a large number of its Ser/Thr/Tyr and Lys residues [6], modulating its additional roles in proliferation, apoptosis, telomere protection, transcription, membrane trafficking, iron metabolism, and receptor mediated cell signaling. When using MS detection, unless specifically searched for in the database when comparing the experimental with the theoretical peptide fragmentation data, PTMs are completely missed. Even if identified, a straightforward quantitative correlation that would enable the summing of contributions of PTM-modified and non-modified peptides is hard to establish. Therefore, peptides that are affected by the presence of PTMs should not be used for normalization in quantitative analysis. Moreover, such PTMs on epitope sites can hinder antigen-antibody interactions and affect the results of western blot analysis

used for data validation, further contributing to the misinterpretation of data (**Figure 3**). Unfortunately, the great majority of proteins carry multiple PTMs that affect multiple amino acids per protein, most commonly including phosphorylation (Ser, Thr, Tyr), acetylation (Lys, Met, Glu, Asp), methylation (His, Lys) and oxidation (Met), and render this selection process extremely difficult. As shown in a sequence alignment of tubulin and actin isoforms (**Table 2**), even the housekeeping proteins most commonly used for normalization may carry an abundance of PTMs. Highlighted in the table are the phosphorylation sites confirmed by 5 or more references, according to the present state of knowledge reflected in the Phosphosite database.

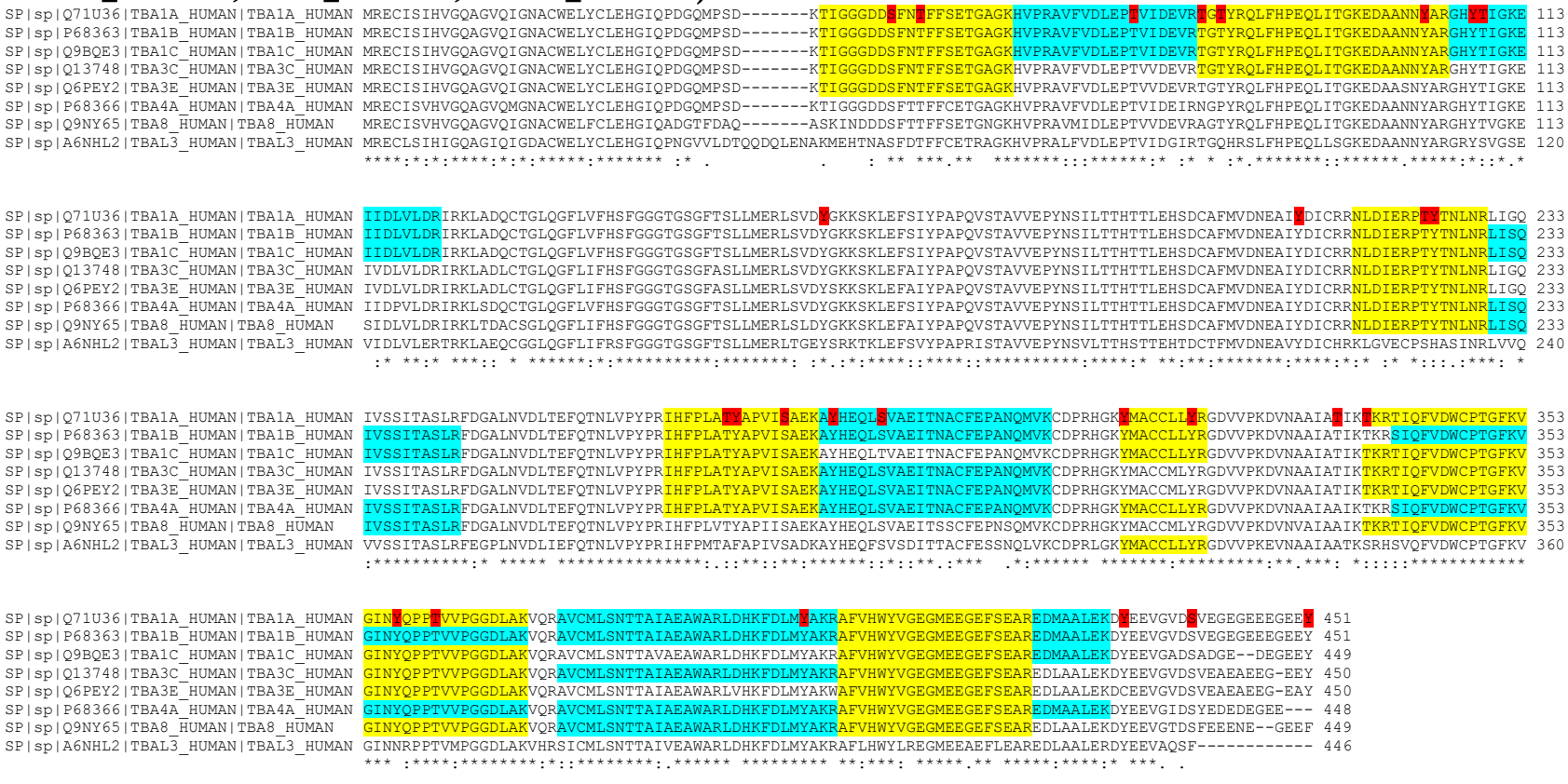


In western blotting experiments, housekeeping proteins are used as loading control. Such experiments rely on specific interactions between an antigen and antibody by a three-dimensional recognition process. When the PTMs affect the three-dimensional structure of the protein or the epitope binding site, this process could be hindered by the PTMs of the target protein. Since PTMs play an important role in signal transduction, some antibodies were developed to detect the modified form only, for example the phosphorylated form of the target protein.

Figure 3. Western blot results can be affected by the presence of PTMs.

Table 2. Sequence alignment of actin, alpha- and beta-tubulins. Highlighted peptides are common to several protein isoforms. Highlighted amino acids (Ser, Thr, Tyr) carry phosphorylation, as confirmed by 5 or more references (*www.phosphosite.org).*

Tubulin alpha (TBA1A_HUMAN, TBA1B_HUMAN, TBA1C_HUMAN, TBA3C_HUMAN, TBA3E_HUMAN, TBA4A_HUMAN, TBA8_HUMAN, TBAL3_HUMAN)



Tubulin beta (TBB1_HUMAN, TBB2A_HUMAN, TBB2B_HUMAN, TBB3_HUMAN, TBB4A_HUMAN, TBB4B_HUMAN, TBB5_HUMAN, TBB6_HUMAN, TBB8_HUMAN)

```
SP|sp|Q9H4B7|TBB1_HUMAN|TBB1_HUMAN   MREIVHIQIGQCGNQIGAKFWEMIGEEHGIDLAGSDRGASALQLERISVYYNEAYGRKYVPRAVLVDLEPGTMDSIRSSKLGALFQPDSFVHGNSGAGNNWAKGHYTEGAELIENVLEVV 120
SP|sp|Q13885|TBB2A_HUMAN|TBB2A_HUMAN MREIVHIQAGQCGNQIGAKFWEVISDEHGIDPTGSYHGDSDLQLERINVYYNEAAGNKYVPRAILVDLEPGTMDSVRSGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDSVLDVV 120
SP|sp|Q9BVA1|TBB2B_HUMAN|TBB2B_HUMAN MREIVHIQAGQCGNQIGAKFWEVISDEHGIDPTGSYHGDSDLQLERINVYYNEATGNKYVPRAILVDLEPGTMDSVRSGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDSVLDVV 120
SP|sp|Q13509|TBB3_HUMAN|TBB3_HUMAN   MREIVHIQAGQCGNQIGAKFWEVISDEHGIDPSGNYVGDSDLQLERISVYYNEASSHKYVPRAILVDLEPGTMDSVRSGAFGHLFRPDNFIFGQSGAGNNWAKGHYTEGAELVDSVLDVV 120
SP|sp|P04350|TBB4A_HUMAN|TBB4A_HUMAN MREIVHLQAGQCGNQIGAKFWEVISDEHGIDPTGTYHGDSDLQLERINVYYNEATGGNYVPRAVLVDLEPGTMDSVRSGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDAVLDVV 120
SP|sp|P68371|TBB4B_HUMAN|TBB4B_HUMAN MREIVHLQAGQCGNQIGAKFWEVISDEHGIDPTGTYHGDSDLQLERINVYYNEATGGKYVPRAVLVDLEPGTMDSVRSGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDSVLDVV 120
SP|sp|P07437|TBB5_HUMAN|TBB5_HUMAN   MREIVHIQAGQCGNQIGAKFWEVISDEHGIDPTGTYHGDSDLQLDRISVYYNEATGGKYVPRAILVDLEPGTMDSVRSGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDSVLDVV 120
SP|sp|Q9BUF5|TBB6_HUMAN|TBB6_HUMAN   MREIVHIQAGQCGNQIGTKFWEVISDEHGIDPAGGYVGDSALQLERINVYYNESSSQKYVPRAALVDLEPGTMDSVRSGPFGQLFRPDNFIFGQTGAGNNWAKGHYTEGAELVDAVLDVV 120
SP|sp|Q3ZCM7|TBB8_HUMAN|TBB8_HUMAN   MREIVLTQIGQCGNQIGAKFWEVISDEHAIDSAGTYHGDSHLQLERINVYYNEASGGRYVPRAVLVDLEPGTMDSVRSGPFGQVFRPDNFIFGQCGAGNNWAKGHYTEGAELMESVMDVV 120
                             *****  * ********:****:*.:**.** :*   * * ***:**.*****: . .***** ***********:**. :* :*:**.*:.*: *****************:: *::**

SP|sp|Q9H4B7|TBB1_HUMAN|TBB1_HUMAN   RHESESCDCLQGFQIVHSLGGGTGSGMGTLLMNKIREEYPDRIMNSFSVMPSPKVSDTVVEPYNAVLSIHQLIENADACFCIDNEALYDICFRTLKLTTPTYGDLNHLVSLTMSGITTSL 240
SP|sp|Q13885|TBB2A_HUMAN|TBB2A_HUMAN RKESESCDCLQGFQLTHSLGGGTGSGMGTLLISKIREEYPDRIMNTFSVMPSPKVSDTVVEPYNATLSVHQLVENTDETYSIDNEALYDICFRTLKLTTPTYGDLNHLVSATMSGVTTCL 240
SP|sp|Q9BVA1|TBB2B_HUMAN|TBB2B_HUMAN RKESESCDCLQGFQLTHSLGGGTGSGMGTLLISKIREEYPDRIMNTFSVMPSPKVSDTVVEPYNATLSVHQLVENTDETYCIDNEALYDICFRTLKLTTPTYGDLNHLVSATMSGVTTCL 240
SP|sp|Q13509|TBB3_HUMAN|TBB3_HUMAN   RKECENCDCLQGFQLTHSLGGGTGSGMGTLLISKVREEYPDRIMNTFSVVPSPKVSDTVVEPYNATLSIHQLVENTDETYCIDNEALYDICFRTLKLATPTYGDLNHLVSATMSGVTTSL 240
SP|sp|P04350|TBB4A_HUMAN|TBB4A_HUMAN RKEAESCDCLQGFQLTHSLGGGTGSGMGTLLISKIREEFPDRIMNTFSVVPSPKVSDTVVEPYNATLSVHQLVENTDETYCIDNEALYDICFRTLKLTTPTYGDLNHLVSATMSGVTTCL 240
SP|sp|P68371|TBB4B_HUMAN|TBB4B_HUMAN RKEAESCDCLQGFQLTHSLGGGTGSGMGTLLISKIREEYPDRIMNTFSVVPSPKVSDTVVEPYNATLSVHQLVENTDETYCIDNEALYDICFRTLKLTTPTYGDLNHLVSATMSGVTTCL 240
SP|sp|P07437|TBB5_HUMAN|TBB5_HUMAN   RKEAESCDCLQGFQLTHSLGGGTGSGMGTLLISKIREEYPDRIMNTFSVVPSPKVSDTVVEPYNATLSVHQLVENTDETYCIDNEALYDICFRTLKLTTPTYGDLNHLVSATMSGVTTCL 240
SP|sp|Q9BUF5|TBB6_HUMAN|TBB6_HUMAN   RKECEHCDCLQGFQLTHSLGGGTGSGMGTLLISKIREEFPDRIMNTFSVMPSPKVSDTVVEPYNATLSVHQLVENTDETYCIDNEALYDICFRTLKLTTPTYGDLNHLVSATMSGVTTSL 240
SP|sp|Q3ZCM7|TBB8_HUMAN|TBB8_HUMAN   RKEAESCDCLQGFQLTHSLGGGTGSGMGTLLLSKIREEYPDRIINTFSILPSPKVSDTVVEPYNATLSVHQLIENADETFCIDNEALYDICSKTLKLPTPTYGDLNHLVSATMSGVTTCL 240
                             *:*.* ********:.***************:.*:***:****:*:**::***************.**:***:**:*  :.********** :**** ************ ****:**.*

SP|sp|Q9H4B7|TBB1_HUMAN|TBB1_HUMAN   RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTAQGSQQYRALSVAELTQQMFDARNTMAACDLRRGRYLTVACIFRGKMSTKEVDQQLLSVQTRNSSCFVEWIPNNVKVAVCDIPPRG 360
SP|sp|Q13885|TBB2A_HUMAN|TBB2A_HUMAN RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTSRGSQQYRALTVPELTQQMFDSKNMMAACDPRHGRYLTVAAIFRGRMSMKEVDEQMLNVQNKNSSYFVEWIPNNVKTAVCDIPPRG 360
SP|sp|Q9BVA1|TBB2B_HUMAN|TBB2B_HUMAN RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTSRGSQQYRALTVPELTQQMFDSKNMMAACDPRHGRYLTVAAIFRGRMSMKEVDEQMLNVQNKNSSYFVEWIPNNVKTAVCDIPPRG 360
SP|sp|Q13509|TBB3_HUMAN|TBB3_HUMAN   RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTARGSQQYRALTVPELTQQMFDAKNMMAACDPRHGRYLTVATVFRGRMSMKEVDEQMLAIQSKNSSYFVEWIPNNVKVAVCDIPPRG 360
SP|sp|P04350|TBB4A_HUMAN|TBB4A_HUMAN RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTSRGSQQYRALTVPELTQQMFDAKNMMAACDPRHGRYLTVAAVFRGRMSMKEVDEQMLSVQSKNSSYFVEWIPNNVKTAVCDIPPRG 360
SP|sp|P68371|TBB4B_HUMAN|TBB4B_HUMAN RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTSRGSQQYRALTVPELTQQMFDAKNMMAACDPRHGRYLTVAAVFRGRMSMKEVDEQMLNVQNKNSSYFVEWIPNNVKTAVCDIPPRG 360
SP|sp|P07437|TBB5_HUMAN|TBB5_HUMAN   RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTSRGSQQYRALTVPELTQQVFDAKNMMAACDPRHGRYLTVAAVFRGRMSMKEVDEQMLNVQNKNSSYFVEWIPNNVKTAVCDIPPRG 360
SP|sp|Q9BUF5|TBB6_HUMAN|TBB6_HUMAN   RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTSRGSQQYRALTVPELTQQMFDARNMMAACDPRHGRYLTVATVFRGPMSMKEVDEQMLAIQSKNSSYFVEWIPNNVKVAVCDIPPRG 360
SP|sp|Q3ZCM7|TBB8_HUMAN|TBB8_HUMAN   RFPGQLNADLRKLAVNMVPFPRLHFFMPGFAPLTSRGSQQYRALTVAELTQQMFDAKNMMAACDPRHGRYLTAAAIFRGRMPMREVDEQMFNIQDKNSSYFADWLPNNVKTAVCDIPPRG 360
                             ************************************::********:* *****:**::* ***** *:*****.* :*** *  :***:*:: :* :*** *.:*:******.*********

SP|sp|Q9H4B7|TBB1_HUMAN|TBB1_HUMAN   LSMAATFIGNNTAIQEIFNRVSEHFSAMFKRKAFVHWYTSEGMDINEFGEAENNIHDLVSEYQQFQDAKAVLEEDEEVTEEAEMEPEDKGH 451
SP|sp|Q13885|TBB2A_HUMAN|TBB2A_HUMAN LKMSATFIGNSTAIQELFKRISEQFTAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVSEYQQYQDATADEQGEFEEEEGEDEA------ 445
SP|sp|Q9BVA1|TBB2B_HUMAN|TBB2B_HUMAN LKMSATFIGNSTAIQELFKRISEQFTAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVSEYQQYQDATADEQGEFEEEEGEDEA------ 445
SP|sp|Q13509|TBB3_HUMAN|TBB3_HUMAN   LKMSSTFIGNSTAIQELFKRISEQFTAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVSEYQQYQDATAEEEGEMYEDDEEESEAQGPK- 450
SP|sp|P04350|TBB4A_HUMAN|TBB4A_HUMAN LKMAATFIGNSTAIQELFKRISEQFTAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVSEYQQYQDATAE-EGEFEEEAEEEVA------ 444
SP|sp|P68371|TBB4B_HUMAN|TBB4B_HUMAN LKMSATFIGNSTAIQELFKRISEQFTAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVSEYQQYQDATAEEEGEFEEEAEEEVA------ 445
SP|sp|P07437|TBB5_HUMAN|TBB5_HUMAN   LKMAVTFIGNSTAIQELFKRISEQFTAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVSEYQQYQDATAEEEEDFGEEAEEEA------- 444
SP|sp|Q9BUF5|TBB6_HUMAN|TBB6_HUMAN   LKMASTFIGNSTAIQELFKRISEQFSAMFRRKAFLHWFTGEGMDEMEFTEAESNMNDLVSEYQQYQDATANDGEEAFEDEEEEIDG----- 446
SP|sp|Q3ZCM7|TBB8_HUMAN|TBB8_HUMAN   LKMSATFIGNNTAIQELFKRVSEQFTAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVSEYQQYQDATAEEEEDEEYAEEEVA------- 444
                             *.*: *****.*****:*:*:**:*:***:****:**:*.****  ** ***.*:.********:***.*    :
```

Actin (ACTA_HUMAN, ACTB_HUMAN, ACTC_HUMAN, ACTG_HUMAN, ACTH_HUMAN, ACTS_HUMAN)

```
SP|sp|P62736|ACTA_HUMAN|ACTA_HUMAN MCEEEDSTALVCDNGSGLCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIITNWDDMEKIWHHSFYNELRVAPEEHPTLLTEAPLNPKANREKMTQIM 125
SP|sp|P60709|ACTB_HUMAN|ACTB_HUMAN --MDDDIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMTQIM 123
SP|sp|P68032|ACTC_HUMAN|ACTC_HUMAN MCDDEETTALVCDNGSGLVKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIITNWDDMEKIWHHTFYNELRVAPEEHPTLLTEAPLNPKANREKMTQIM 125
SP|sp|P63261|ACTG_HUMAN|ACTG_HUMAN --MEEEIAALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKMTQIM 123
SP|sp|P63267|ACTH_HUMAN|ACTH_HUMAN MCEE-ETTALVCDNGSGLCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIITNWDDMEKIWHHSFYNELRVAPEEHPTLLTEAPLNPKANREKMTQIM 124
SP|sp|P68133|ACTS_HUMAN|ACTS_HUMAN MCDEDETTALVCDNGSGLVKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIITNWDDMEKIWHHTFYNELRVAPEEHPTLLTEAPLNPKANREKMTQIM 125
                              : : :*** *****: ************************************************************:************:*************.******************


SP|sp|P62736|ACTA_HUMAN|ACTA_HUMAN FETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLDLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEKSYELPDGQVI 250
SP|sp|P60709|ACTB_HUMAN|ACTB_HUMAN FETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSYELPDGQVI 248
SP|sp|P68032|ACTC_HUMAN|ACTC_HUMAN FETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLDLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEKSYELPDGQVI 250
SP|sp|P63261|ACTG_HUMAN|ACTG_HUMAN FETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSYELPDGQVI 248
SP|sp|P63267|ACTH_HUMAN|ACTH_HUMAN FETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLDLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEKSYELPDGQVI 249
SP|sp|P68133|ACTS_HUMAN|ACTS_HUMAN FETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLDLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEKSYELPDGQVI 250
                              *****.**********************:********.*************:************************.************************:**********************


SP|sp|P62736|ACTA_HUMAN|ACTA_HUMAN TIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDEAGPSIVHRKCF 377
SP|sp|P60709|ACTB_HUMAN|ACTB_HUMAN TIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF 375
SP|sp|P68032|ACTC_HUMAN|ACTC_HUMAN TIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDEAGPSIVHRKCF 377
SP|sp|P63261|ACTG_HUMAN|ACTG_HUMAN TIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF 375
SP|sp|P63267|ACTH_HUMAN|ACTH_HUMAN TIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKPEYDEAGPSIVHRKCF 376
SP|sp|P68133|ACTS_HUMAN|ACTS_HUMAN TIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNVMSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWITKQEYDEAGPSIVHRKCF 377
                              ***********:******:****.******:********:*********.*:************************************************:* ****:**********
```

(d) The spectral counts for the protein chosen for normalization should be within the linear dynamic range of the detector's response. The spectral count for a protein is proportional to the protein length, the number of its proteotypic peptides, and to its abundance. In the case of a data dependent MS analysis, the spectral count increases with concentration until all proteotypic peptides are detected, and then can increase only with an increase in the chromatographic peak width, if the peptides are still observable in the mass spectrum after expiration of the time the peptides spent on the exclusion list. At this level, the spectral counts change at a different rate, and may not reflect a proportional change in protein/peptide expression levels. For example, in the analysis of a cell extract that resulted in the detection of 985 proteins matched by a total of 4878 peptide spectral counts, the top 2.5-3.0 % most abundant proteins accounted for ~25 % of the total spectral counts, at a level of ~20-100 counts per protein. In contrast, for low abundance proteins, when the peptides are barely detectable (1-2 counts per protein), there is a better proportionality between spectral counts and abundance, but the variability of spectral count data is too high and can lead to a biased interpretation of results if a sufficient number of replicate analyses are not performed**.**

(e) Proteins that generate a reasonable number of unique peptides after proteolytic digestion, rather than shared peptides with other protein isoforms, are preferred. During data processing, the MS data analysis software will search a protein database and will attempt to match the identified peptides to the parent proteins that carry that specific peptide amino acid sequence. This process causes a so-called "protein inference problem," as after proteolytic digestion the connectivity between proteins and peptides is lost [7]. If due to sequence homology a peptide can be matched to more than one parent protein in the database, the actual parent protein cannot be specified with certainty (**Figure 4**). Such a peptide is called shared, non-unique or degenerate peptide. Conventional housekeeping proteins that are used for normalization in biological experiments have several isoforms, and pose, therefore, problems, in terms of identifying the correct parent protein by MS analysis. For example, a protein sequence alignment of 6 actin, 8 alpha tubulin and 9 beta tubulin isoforms indicates 92.1 %, 65.9 % and 69.4 % sequence homology, respectively (**Table 3**). A simple approach to address this challenge would be to remove the ambiguity

by simply ignoring the shared peptide from the dataset [8]. Another approach would involve distributing the count of the shared peptide among the parent proteins in proportion to the total spectral counts associated with each contributing parent protein [9]. Such approaches can be implemented, however, only if there exist several other unique peptides that could be used to confidently identify the parent protein of interest. This is certainly not the case of actin, and given that in most large-scale MS experiments the great majority of proteins are identified by only very few peptides, even proteins such as tubulin cannot be uniquely identified. Taking into account, however, that many protein isoforms perform identical or similar functions, the shared peptide problem could be addressed by considering the set of isoforms as a whole set, and using for comparison and normalization the sum of all peptide contributions to the set. Overall, while not trivial, a prudent comparison of the experimental results with protein sequence and PTM databases, may enable the selection of PTM-free peptides that are representative of unique proteins that could be used for normalization [10].

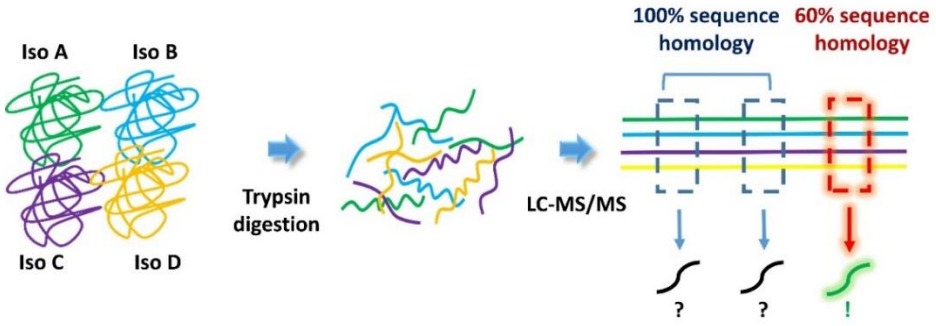


Figure 4. Shared peptides from homologous proteins.

Table 3. Sequence homology among the isoforms of actin and tubulin.

Protein	Sequence homology	PTMs
Actin ACTA_HUMAN, ACTB_HUMAN, ACTC_HUMAN, ACTG_HUMAN, ACTH_HUMAN, ACTS_HUMAN	92.1 %	Methylation Acetylation Oxidation Ubl conjugation
Tubulin alpha TBA1A_HUMAN, TBA1B_HUMAN, TBA1C_HUMAN, TBA3C_HUMAN, TBA3E_HUMAN, TBA4A_HUMAN, TBA8_HUMAN, TBAL3_HUMAN	65.9 %	Acetylation Nitration Phosphorylation Isopeptide bond Methylation Ubl conjugation
Tubulin beta TBB1_HUMAN, TBB2A_HUMAN, TBB2B_HUMAN, TBB3_HUMAN, TBB4A_HUMAN, TBB4B_HUMAN, TBB5_HUMAN, TBB6_HUMAN, TBB8_HUMAN	69.4 %	Acetylation Isopeptide bond Methylation Phosphorylation Ubl conjugation

4.2 Proposed protein set for normalization of spectral count data generated by MS analysis of cell extracts.

To identify a representative set of proteins for MS spectral count data normalization, MCF-7 and MCF-10A cells were cultured in appropriate growth media, arrested in the cell cycle by serum deprivation, and released with medium containing hormones or growth factors, respectively. The percent of G1:S:G2/M cells in the different stages of cell cycle was 80:10:7 in arrested cells and 28:60:10 in released cells (CV = 2-12 %). The cell extracts were separated into nuclear and cytoplasmic fractions. This process generated two complementary cell fractions (nuclear and cytoplasmic), in two complementary stages of the cell cycle (non-proliferating G1 and proliferating S), from two functionally distinct cell lines (cancerous and non-tumorigenic). The cell extracts were processed and prepared for LC-MS analysis. Data-dependent MS analysis resulted typically in the identification of 800-1000 proteins per LC-MS run (FDR<3 % at the protein and <1 % at the peptide level, respectively), matched by 4000-4200 spectral counts. Multiple tandem MS spectra per peptide were allowed in the dataset. Five LC-MS technical replicates were performed for each cell fraction to increase the number and confidence in the spectral count data, and

three biological replicates were processed to enable the evaluation of statistical significance. This experimental approach resulted in the identification of a total of 3700 proteins. The LC-MS technical replicates were averaged, and the data were normalized based on a grand average calculated from the total spectral counts corresponding to the 12 nuclear and 12 cytoplasmic fractions, respectively (2 cell lines x 2 cell cycle stages x 3 biological replicates). Under the underlying hypothesis that the expression level of an ideal endogenous protein suitable for normalization will not change in response to a major biological perturbation such as a change in cell cycle stage, or a transition from a non-cancerous to a cancerous cell state, spectral count CV values for each protein in the complementary nuclear and cytoplasmic fractions were calculated separately, and used to sort the two lists to determine the proteins that exhibited the smallest variations in spectral counts across the 12 fractions. These proteins represent the best candidates for normalization and validation of differential expression data. **Tables 4A** and **B** provide a set of **103** proteins (34 nuclear and 75 cytoplasmic), their average count in the nuclear/cytoplasmic fraction, the standard deviation, the associated coefficients of variation, the cellular location and the associated PTMs. These proteins were selected from the list of 3700, according to the following criteria: (a) the average number of matching spectral counts in the 12 cell states had to be higher than 4 to avoid variability concerns at the low-end of the spectral count range, and less than 40-50 to avoid saturation effects at the high-end of the range; proteins with much larger spectral count did not qualify, in fact, for selection, except PRKDC (DNA-dependent protein kinase catalytic subunit) in the nuclear fraction and KPYM (pyruvate kinase isozymes M1/M2) in the cytoplasmic fraction; and (b) the reproducibility of protein identifications in a particular cell fraction, i.e., nuclear or cytoplasmic, had to be reflected by a CV value of less than 30 %. The actin, tubulin, GAPDH (protein name G3P) and 6 other proteins (Ku70, Ku86, nucleolin, HSP72, calmodulin and peptidyl-prolyl cis-trans isomerase) were common to both nuclear and cytoplasmic fractions (**Tables 4C** and **D**).

The expected cellular location and biological function of the proteins was assigned by using bioinformatics tools enabled by the David, STRING, Genecards and Uniprot websites. The cytoplasmic proteins were involved in biological processes encompassing primarily

apoptosis, cellular redox, carbohydrate/nucleotide and various other metabolic processes, protein folding, transport and degradation, translation, and cell cycle/signaling. The location of these proteins was assigned mainly to the cytoplasm, but also to the mitochondria, ER, Golgi, proteasome, and to a lesser extent to the nucleus/nucleoplasm and nuclear envelope. The nuclear proteins were involved in processes encompassing mRNA processing and metabolism, DNA repair and metabolism, chromosome/telomere organization and maintenance, and cell cycle/signaling. Their cellular location was assigned to the nuclear lumen, nucleolus, nucleoplasm, chromosome, nuclear membrane, spliceosome, ribonucleoprotein complex, ER, and to a lesser extent to the cytoskeleton/cytosol. Overall, the location and functional roles associated with the great majority of the selected proteins confirms that these proteins perform mainly routine housekeeping operations, and that their selection for normalization and validation functions is well justified.

The presence of certain proteins in an unexpected fraction was, however, observed. As the proteins with the largest spectral counts were identifiable only in one cellular fraction but not in the other (PRKDC in the nuclear, and KPYM in the cytoplasmic), simple cross-contamination was assumed to be minimal, and below the limit of detection. Therefore, alternative explanations were sought. Nuclear proteins such as Ku70, Ku86 are associated with the chromosomes, and their presence in the cytoplasmic fraction can be rationalized through the contribution of G2/M cells to both G1 and S-phase cell batches (~7-10 %). Nucleolin is a nuclear protein, but can be localized in the mRNP (messenger ribonuclear protein) granules that contain untranslated mRNA. mRNAs are coated with proteins and form mRNP complexes that enter the cytoplasm and engage into translation, or remain translationally inactive and assemble as cytoplasmic mRNP granules. The peptidyl-prolyl cis-trans isomerase protein is localized to the ER, therefore it is not surprising that could be identified in both fractions, as the complete separation of the ER from the nucleus is difficult to accomplish during the fractionation of the nuclear and cytoplasmic cell fractions. It has roles in protein folding and catalyzes the cis-trans isomerization of proline imidic peptide bonds in oligopeptides. Interestingly, the other cytoplasmic proteins that contaminated the nuclear fraction, each have some role in the mitotic process, being

associated with, or binding to the chromosomes. For example, the tubulins are the major components of the mitotic spindle apparatus, while calmodulin (a Ca-binding/phosphorylase kinase) is localized during mitosis to the spindle poles and the spindle microtubules. HSP72 is a molecular chaperone that mediates the folding of newly translated proteins in the cytosol or within organelles, and is involved in G2/M-specific positive regulation of cyclin-dependent protein kinase activity. Through inference, is believed to be part of the synaptonemal complex that forms between homologous chromosomes to mediate chromosome pairing, synapsis and recombination during meiosis.

Additional proteins identified in the nuclear fraction, but otherwise known to have cytoplasmic localization, included nuclear ribonucleoproteins associated with the ER (HNRPQ, HNRPD), septin associated with the spindle/chromosome/kinetochore, and serine/threonine-protein phosphatase PP1 involved in signaling. Cytoplasmic proteins associated with the cell membrane (that did not break apart during the mild lysis conditions under which the cell nuclei were separated from the cytoplasm), cell cortex and actin cytoskeleton such as the Ras GTPase-activating-like protein IQGAP1, filamin, ezrin, and spectrin, are believed to be contaminants that were deposited together with the cell membrane into the nuclear fraction during centrifugation and separation of the cytoplasmic from the nuclear fraction. Although in our cell cycle experiments these proteins displayed stable spectral count data, their use for normalization should be exercised with caution, either due to the above artifacts associated with the cell fractionation process, or due to their multiple functional roles that may lead to relocation in various cell compartments. For example, spectrins are a multifunctional actin-binding protein family and major components of the membrane cytoskeletal proteins. βII-spectrin, one of the spectrin isoform, is involved in a variety of biological processes that affect cell morphology, mechanical properties, compaction and accumulation of E-cadherin in the epithelial cell-cell contact. βII-spectrin plays a role in protein and phospholipid delivery, and cell cycle regulation in TGF-β signaling. Although it is mainly located in cytoplasm, βII-spectrin can be also localized in the nucleus as a component of the nucleoskeleton. Its distribution and expression level varies by disease state and cell type [11]. A spectrin-like protein, with similar properties to a structural protein, was also found to be expressed in the nucleus of

a plant (*Allium cepa*) and suggested that it could be involved in multiple nuclear functions such as structural component, DNA repair, transcription and RNA processing [12].

The protein that showed the lowest CV value in the nuclear fraction, MCTS1, is expressed in almost all tissues except the lung, and was found to be overexpressed in a variety of T-cell and B-cell lymphomas [13]. In NIH 3T3 fibroblasts, MCTS1 performs several functions such as decreasing cell doubling time and decreasing the duration of the G1 transit time and/or G1-S transition. In T-cells, MCTS1 is primarily localized in the cytoplasm, and its levels were found to be invariant during the various phases of the cell cycle. In MCF-7 breast cancer cells, on the other hand, after overexpression, MCTS1 was predominantly localized in the nucleus instead of the cytoplasm [14].

4.3 Assessment of the proposed protein set.

The proposed protein set is expected to be best suited for validating large scale protein differential expression data. Under identical experimental conditions, the same amount of protein subjected to analysis should generate roughly the same number of total spectral counts that match the protein set identified in the analysis. In the case of replicate analyses, the average of the total spectral counts can be used to normalize the data and reduce the impact of random experimental errors. In our experiments, a set of 800-1000 proteins was typically matched by 4000-4200 spectral counts (CV<7-8 %). Based on the change in number of counts pertaining to a set of protein spikes that were used to evaluate experimental variability (**Table 5**), the threshold for qualifying a protein as differentially expressed was set to a minimum of a two-fold change in spectral counts [15]. Once proteins that pass such a threshold can be identified in an experiment, differential expression must be validated by alternative analysis methods, such as western blotting or complementary MS technique such as MRM. The use of external standards is a very effective approach for evaluating the reproducibility and variability of an experimental procedure. Unlike with internal standards (i.e., endogenous proteins), the amount, nature and time when the external standard is introduced in the sample can be manipulated, allowing researchers to monitor the performance of particular experimental steps. On the other hand, endogenous proteins can be successfully used to assess whether the observed changes during a

biological experiment are trustworthy or not under the impact of biological variability. As observed from **Tables 4** and **5**, the spectral count CVs associated with the proposed protein set (<30 %) were less or equal to the CVs associated with the standard protein spikes, denoting that: (a) the expression level of the proposed proteins was constant during the experiment, and (b) the biological variability did not exceed the levels of experimental variability. To ultimately assess whether the proposed protein set can be used as a base-line control for validating quantitative data, the set was evaluated against the cytoplasmic fraction of a batch of SKBR3 cells with a percent cell cycle distribution G1:S:G2/M of 72:18:10. The experiments were conducted 4 years apart, using completely different reagent batches and instrumentation supplies. The average spectral count data in SKBR3, and the percent change vs. MCF-7/MCF-10 are shown in the right-end columns of **Table 4B**. The percent change in the cytoplasmic set was <50 % for 59 proteins, between 50 and 84 % for an additional 12 proteins, and in excess of 100 % for two proteins. The small number of spectral counts for the 12 proteins, and the lack of replicates in SKBR3 data, are thought to be the primary factors that contributed to the larger than expected variability in the detection of these proteins. We expect, however, to be able to validate this additional sub-set once replicate data will become available for SKBR3. Despite multiple functional roles and the presence of PTMs, the rest of 59 proteins, displayed changes mostly within the variability limits encountered in MCF-7/MCF-10. Commonly used actin and tubulin were also well-behaved, the tubulin isoforms presenting the most stable distribution of spectral counts among all cell states and fractions. The source of large variation in the GAPDH counts (the largest in the dataset), on the other hand, is challenging its broad utilization for validation of biological quantitative data, and requires further investigation. The summing of all spectral counts for a particular category resulted in a sizeable reduction in CV values, and may represent a more effective approach for data normalization or validation. For example, for the entire set of proposed proteins, individual CV values of 10-30 % could be reduced to 5-6 % by such a summing process. Similar trends were observable for the actin and tubulin isoforms, as well. Such a summing process should be used, however, with caution, only to minimize the random noise contributions associated with proteins with known, expected behavior. Assuming that the differential expression threshold is maintained at the 2-fold change level, overall, the results confirm the

applicability of the proposed proteins set for validation of MS spectral count data. While the combination of the two sets of reference proteins, endogenous and external standards, will ensure a rigorous assessment of large-scale protein expression results, smaller or larger protein sub-sets could prove to be better suited for particular biological applications. For biological experiments which induce known changes in protein expression, additional protein sets which change their spectral counts by an anticipated factor could be further developed for confirming the expected outcomes.

4.4 Nuclear/cytoplasmic markers.

As the two nuclear and cytoplasmic protein sets were selected from the list of 3700 proteins solely based on minimal variability in spectral counts, and as only 6 proteins from the entire set were identifiable in both cellular fractions, it is expected that some of these proteins could represent useful organelle markers, as well. Based on biological function, the best suited marker proteins included two DNA damage repair proteins in the nuclear fraction (PRKDC and PARP1), and the peroxiredoxins (PRDX1, 2, and 6) and the mitochondrial proteins in the cytoplasmic fraction. An interesting protein was KPYM_HUMAN Pyruvate kinase isozymes M1/M2 which was identified in the cytoplasmic fraction. This protein is known to be a cytoplasmic protein, but localizes to the nucleus in response to apoptotic stimuli. For this dataset that was generated from G1-arrested cells, after 48 h of serum deprivation, this protein represented a reassuring confirmation of cell viability and adequate experimental design.

Table 4A. Proposed protein set for data normalization and validation in the nuclear fractions.

Swissprot	Protein name	MW	Avg	SD	CV	Localization	PTMs
Q502X6	Q502X6_HUMAN MCTS1 protein	19216.0	9.4	1.1	11.3	Cytoplasm (Nuclear relocalization after DNA damage)	Phosphoprotein
P78527	PRKDC_HUMAN DNA-dependent protein kinase catalytic subunit	468786.9	53.0	6.3	12.0	Nucleus	Acetylation, Phosphoprotein, S-nitrosylation
Q9UHD8	SEPT9_HUMAN Septin-9	65360.9	5.6	0.7	13.1	Cytoskeleton	Acetylation, Phosphoprotein
Q9BZZ5	API5_HUMAN Apoptosis inhibitor 5	57525.2	4.9	0.7	14.3	Mainly nuclear. Can also be cytoplasmic	Acetylation, Phosphoprotein
Q92499	DDX1_HUMAN ATP-dependent RNA helicase DDX1	82379.9	9.2	1.4	15.3	Nucleus, Cytoplasm	Acetylation, Phosphoprotein
P62136	PP1A_HUMAN Serine/threonine-protein phosphatase PP1-alpha catalytic subunit	37487.8	8.6	1.5	17.3	Cytoplasm, Nucleoplasm, Nucleolus	Acetylation, Phosphoprotein
P62314	SMD1_HUMAN Small nuclear ribonucleoprotein Sm D1	13273.4	7.5	1.3	17.6	Nucleus	Methylation

P54652	HSP72_HUMAN Heat shock-related 70 kDa protein 2	69978.0	5.9	1.0	17.7	Synaptonemal complex, male germ cell nucleus, mitochondrion, cell surface, CatSper complex	Methylation, Phosphoprotein
Q15637	SF01_HUMAN Splicing factor 1	68286.1	6.5	1.2	18.2	Nucleus	Acetylation, Phosphoprotein
Q92841	DDX17_HUMAN Probable ATP-dependent RNA helicase DDX17	72326.0	15.2	2.8	18.3	Nucleus, Nucleolus	Isopeptide bond, Phosphoprotein, Ubl conjugation
Q14103	HNRPD_HUMAN Heterogeneous nuclear ribonucleoprotein D0	38410.3	6.2	1.1	18.4	Nucleus, Cytoplasm	Acetylation, Methylation, Phosphoprotein
P62158	CALM_HUMAN Calmodulin	16826.8	5.4	1.0	18.6	Spindle, Spindle pole	Acetylation, Isopeptide bond, Methylation, Phosphoprotein, Ubl conjugation
Q6IBH5	Q6IBH5_HUMAN Peptidyl-prolyl cis-trans isomerase	23713.5	14.8	2.8	18.6	ER lumen, Melanosome	Glycoprotein
P12956	KU70_HUMAN ATP-dependent DNA helicase 2 subunit 1	69799.2	14.8	2.8	18.8	Nucleus, Chromosome	Acetylation, Phosphoprotein
P26599	PTBP1_HUMAN Polypyrimidine	57185.8	22.5	4.3	19.1	Nucleus	Acetylation, Phosphoprotein

	tract-binding protein 1						
P19338	NUCL_HUMAN Nucleolin	76568.5	26.6	5.1	19.3	Nucleolus	Acetylation, Methylation, Phosphoprotein
P46940	IQGA1_HUMAN Ras GTPase-activating-like protein IQGAP1	189132.9	27.6	5.6	20.4	Cell membrane	Acetylation, Phosphoprotein
Q16630	CPSF6_HUMAN Cleavage and polyadenylation specificity factor subunit 6	59173.5	5.4	1.1	20.6	Nucleus	Phosphoprotein
O60506	HNRPQ_HUMAN Heterogeneous nuclear ribonucleoprotein Q	69559.6	16.2	3.6	22.2	Cytoplasm	Acetylation, Isopeptide bond, Methylation, Phosphoprotein, Ubl conjugation
Q13813	SPTA2_HUMAN Spectrin alpha chain, brain	284362.5	43.3	9.6	22.3	Cytoskeleton, Cell cortex	Acetylation, Phosphoprotein
Q9Y3I0	CV028_HUMAN UPF0027 protein C22orf28	55175.0	7.2	1.6	22.3	Cytoplasm	
P60660	MYL6_HUMAN Myosin light polypeptide 6	16919.1	8.8	2.0	22.3		Acetylation, Phosphoprotein

P24534	EF1B_HUMAN Elongation factor 1-beta	24748.3	4.6	1.0	22.6		Acetylation, Phosphoprotein
P62841	RS15_HUMAN 40S ribosomal protein S15	17029.2	9.3	2.1	22.7		Acetylation
Q15717	ELAV1_HUMAN ELAV-like protein 1	36069.2	6.6	1.5	22.9	Cytoplasm, Nucleus	Acetylation, Methylation, Phosphoprotein
P21333	FLNA_HUMAN Filamin-A	280561.4	78.7	18.1	23.1	Cell cortex, Cytoskeleton.	Acetylation, Phosphoprotein
O00571	DDX3X_HUMAN ATP-dependent RNA helicase DDX3X	73198.1	13.0	3.2	24.2	Nucleus speckle, Cytoplasm, Mitochondrion outer membrane	Acetylation, Phosphoprotein
Q9UMS4	PRP19_HUMAN Pre-mRNA-processing factor 19	55146.4	6.0	1.5	24.5	Nucleoplasm, Spindle	Acetylation
P15311	EZRI_HUMAN Ezrin	69369.8	15.4	3.8	24.6	Apical and Peripheral membrane	Acetylation, Phosphoprotein
Q15019	SEPT2_HUMAN Septin-2	41461.3	7.3	1.8	25.0	Cytoskeleton., Spindle, Kinetochore.	Acetylation, Phosphoprotein
Q15233	NONO_HUMAN Non-POU domain-containing octamer-binding protein	54197.4	28.2	7.1	25.2	Nucleolus, Nucleus speckle	Acetylation, Phosphoprotein
P13010	KU86_HUMAN ATP-dependent	82652.4	16.1	4.4	27.4	Nucleolus, Chromosome	Acetylation, Phosphoprotein

	DNA helicase 2 subunit 2						
P27695	APEX1_HUMAN DNA-(apurinic or apyrimidinic site) lyase	35532.2	5.4	1.5	28.6	Nucleolus, Nucleus speckle, ER	Acetylation, Cleavage on pair of basic residues, Disulfide bond, Nitration, Phosphoprotein, S-nitrosylation, Ubl conjugation
P09874	PARP1_HUMAN Poly [ADP-ribose] polymerase 1	113012.4	19.5	5.9	30.0	Nucleolus	ADP-ribosylation, Acetylation, Phosphoprotein, S-nitrosylation
	Sum		**534.3**	**26.3**	**4.9**		

Table 4B. Proposed protein set for data normalization and validation in the cytoplasmic fractions.

Swissprot	Protein name	MW	Avg	SD	CV	Localization	PTMs	Avg SKBR3	% Change
P06576	ATPB_HUMAN ATP synthase subunit beta, mitochondrial precursor	56524.7	28.0	2.6	9.2	Mitochondrion inner membrane	Acetylation, Phosphoprotein	22.0	21.1
P07741	APT_HUMAN Adenine phosphoribosyltransferase	19595.4	6.2	0.6	9.3	Cytoplasm	Acetylation, Phosphoprotein	1.0	83.9

P62158	CALM_HUMAN Calmodulin	16826.8	7.5	0.8	10.6	Spindle pole	Acetylation, Isopeptide bond, Methylation, Phosphoprotein, Ubl conjugation	5.9	20.3
Q2KHP4	Q2KHP4_HUMAN HSPA5 protein	72377.6	31.6	3.5	10.9			37.1	16.0
P00558	PGK1_HUMAN Phosphoglycerate kinase 1	44586.2	27.3	3.0	11.1	Cytoplasm	Acetylation, Phosphoprotein	20.4	25.7
P60900	PSA6_HUMAN Proteasome subunit alpha type-6	27381.8	3.5	0.4	11.2	Cytoplasm, Nucleus	Acetylation, Glycoprotein, Isopeptide bond, Phosphoprotein, Ubl conjugation	2.1	40.2
P22314	UBA1_HUMAN Ubiquitin-like modifier-activating enzyme 1	117774.5	21.1	2.6	12.5		ISGylated		
Q05524	ENO1B_HUMAN Alpha-enolase, lung specific	49446.4	10.6	1.4	13.4			7.8	28.2
P25705	ATPA_HUMAN ATP synthase subunit alpha, mitochondrial precursor	59713.7	17.0	2.3	13.5	Mitochondrion inner membrane, Peripheral membrane	Acetylation, Phosphoprotein, Pyrrolidone carboxylic acid	14.1	16.3
P15880	RS2_HUMAN 40S ribosomal protein S2	31304.6	5.5	0.8	14.6		Acetylation, Citrullination, Methylation, Phosphoprotein	3.9	30.1

P14618	KPYM_HUMAN Pyruvate kinase isozymes M1/M2	57900.2	103.9	15.3	14.7	Cytoplasm, Nucleus	Acetylation, Hydroxylation, Phosphoprotein, Ubl conjugation	95.6	7.6
P50395	GDIB_HUMAN Rab GDP dissociation inhibitor beta	50630.9	8.8	1.4	15.4	Cytoplasm, Peripheral membrane	Acetylation, Phosphoprotein	10.0	14.9
O75947	ATP5H_HUMAN ATP synthase subunit d, mitochondrial	18479.5	5.4	0.9	15.8	Mitochondrion inner membrane	Acetylation	2.6	51.9
P06744	G6PI_HUMAN Glucose-6-phosphate isomerase	63107.3	19.2	3.1	16.3	Cytoplasm, Secreted	Acetylation, Phosphoprotein, Ubl conjugation	17.6	9.9
Q14974	IMB1_HUMAN Importin subunit beta-1	97108.2	11.5	1.9	16.5	Cytoplasm, Nucleus envelope	ADP-ribosylation, Acetylation, Isopeptide bond, Phosphoprotein, Ubl conjugation	5.3	53.9
P50454	SERPH_HUMAN Serpin H1 precursor	46411.3	5.5	0.9	16.6	ER lumen	Glycoprotein		
P48047	ATPO_HUMAN ATP synthase subunit O, mitochondrial precursor	23262.7	5.3	0.9	16.7	Mitochondrion inner membrane	Acetylation	4.1	21.6
P13804	ETFA_HUMAN Electron transfer flavoprotein subunit alpha, mitochondrial	35057.6	7.1	1.2	16.9	Mitochondrion matrix	Acetylation	2.1	72.3

P60174	TPIS_HUMAN Triosephosphate isomerase	26652.7	17.2	2.9	16.9		Acetylation, Nitration, Phosphoprotein	15.3	9.4
P29401	TKT_HUMAN Transketolase	67834.9	12.2	2.1	16.9		Acetylation, Phosphoprotein	10.8	11.9
Q13200	PSMD2_HUMAN 26S proteasome non-ATPase regulatory subunit 2	100136.0	6.5	1.1	17.0		Acetylation, Phosphoprotein	2.2	64.4
P06748	NPM_HUMAN Nucleophosmin	32554.9	7.1	1.2	17.4	Nucleolus, Nucleoplasm, Centrosome	ADP-ribosylation, Acetylation, Disulfide bond, Isopeptide bond, Phosphoprotein, Ubl conjugation	6.3	11.8
P63241	IF5A1_HUMAN Eukaryotic translation initiation factor 5A-1	16821.4	8.5	1.5	17.5	Cytoplasm, Nucleus, ER membrane, Peripheral membrane, Nuclear pore complex	Acetylation, Hypusine	3.7	55.2
P30101	PDIA3_HUMAN Protein disulfide-isomerase A3 precursor	56746.8	20.4	3.6	17.5	ER lumen, Melanosome	Disulfide bond	12.3	39.8
O00303	EIF3F_HUMAN Eukaryotic translation initiation factor 3 subunit F	37540.2	3.9	0.7	17.6	Cytoplasm	Acetylation, Phosphoprotein	2.0	46.2
Q01105	SET_HUMAN Protein SET	33468.7	6.0	1.1	18.0	Cytosol, ER	Acetylation, Isopeptide bond,	4.5	27.7

							Methylation, Phosphoprotein, Ubl conjugation		
P60981	DEST_HUMAN Destrin	18493.5	4.7	0.8	18.1		Acetylation, Phosphoprotein, Ubl conjugation	4.3	11.8
P09622	DLDH_HUMAN Dihydrolipoyl dehydrogenase, mitochondrial precursor	54116.0	4.6	0.8	18.1	Mitochondrion matrix	Acetylation, Disulfide bond, Phosphoprotein	5.9	25.7
P31943	HNRH1_HUMAN Heterogeneous nuclear ribonucleoprotein H	49198.4	4.0	0.7	18.4	Nucleoplasm	Acetylation, Methylation, Phosphoprotein	6.1	50.7
P62333	PRS10_HUMAN 26S protease regulatory subunit S10B	44145.2	4.4	0.8	18.6	Cytoplasm, Nucleus	Acetylation	2.2	50.4
P54652	HSP72_HUMAN Heat shock-related 70 kDa protein 2	69978.0	7.2	1.3	18.8		Methylation, Phosphoprotein	9.9	39.6
P25787	PSA2_HUMAN Proteasome subunit alpha type-2	25882.3	4.1	0.8	18.8	Cytoplasm, Nucleus	Acetylation, Nitration, Phosphoprotein	3.5	14.9
P10809	CH60_HUMAN 60 kDa heat shock protein, mitochondrial precursor	61016.5	35.9	6.8	19.0	Mitochondrion matrix	Acetylation, Phosphoprotein	26.3	27.5
P04792	HSPB1_HUMAN Heat shock protein beta-1	22768.5	20.9	4.0	19.1	Cytoplasm, Nucleus, Spindle	Acetylation, Phosphoprotein	10.0	52.9
P31946	1433B_HUMAN 14-3-3 protein beta/alpha	28064.8	12.4	2.4	19.2	Cytoplasm, Melanosome	Acetylation, Nitration, Phosphoprotein	12.2	1.2

P27824	CALX_HUMAN Calnexin precursor	67526.0	15.0	2.9	19.3	ER membrane, Melanosome	Acetylation, Disulfide bond, Lipoprotein, Palmitate, Phosphoprotein	15.3	2.8
P40926	MDHM_HUMAN Malate dehydrogenase, mitochondrial precursor	35508.8	18.1	3.5	19.3	Mitochondrion matrix	Acetylation	25.5	39.6
Q15181	IPYR_HUMAN Inorganic pyrophosphatase	32639.2	8.4	1.6	19.3	Cytoplasm	Acetylation, Phosphoprotein	9.4	12.8
P30086	PEBP1_HUMAN Phosphatidylethanolamine-binding protein 1	21043.7	6.1	1.2	19.3	Cytoplasm	Disulfide bond, Phosphoprotein	5.1	18.2
P55072	TERA_HUMAN Transitional endoplasmic reticulum ATPase	89265.9	12.8	2.5	19.7	Cytosol, ER, Nucleus	Acetylation, Methylation, Phosphoprotein, Ubl conjugation	11.6	9.4
P23528	COF1_HUMAN Cofilin-1	18490.7	24.8	4.9	19.8	Nucleus matrix, cytoskeleton, Peripheral membrane	Acetylation, Phosphoprotein	23.4	5.6
P13667	PDIA4_HUMAN Protein disulfide-isomerase A4 precursor	72887.1	7.9	1.6	19.8	ER lumen, Melanosome	Acetylation, Disulfide bond	4.1	50.3
P08107	HSP71_HUMAN Heat shock 70 kDa protein 1	70009.2	18.7	3.7	20.0	Cytoplasm	Acetylation, Methylation, Phosphoprotein	9.6	47.8
P39019	RS19_HUMAN 40S ribosomal protein S19	16050.5	4.5	0.9	20.7	Nucleus	Acetylation	3.5	22.7

P12956	KU70_HUMAN ATP-dependent DNA helicase 2 subunit 1	69799.2	10.8	2.2	20.7	Nucleus, Chromosome	Acetylation, Phosphoprotein	7.5	30.8
Q6IBH5	Q6IBH5_HUMAN Peptidyl-prolyl cis-trans isomerase	23713.5	5.0	1.0	20.8	ER lumen, Melanosome	Glycoprotein	5.5	11.2
Q6S8J3	A26CA_HUMAN ANKRD26-like family C member 1A	121285.6	24.9	5.2	20.9		Isopeptide bond, Ubl conjugation	16.1	36.6
P14625	ENPL_HUMAN Endoplasmin precursor	92411.2	20.2	4.2	21.0	ER lumen, Melanosome	Disulfide bond, Glycoprotein, Phosphoprotein	18.0	11.5
P07737	PROF1_HUMAN Profilin-1	15044.6	23.4	5.0	21.3	Cytoskeleton	Acetylation, Isopeptide bond, Phosphoprotein, Ubl conjugation	17.5	25.9
P60842	IF4A1_HUMAN Eukaryotic initiation factor 4A-I	46124.6	12.0	2.5	21.3		Acetylation, Phosphoprotein	7.1	37.5
P62258	1433E_HUMAN 14-3-3 protein epsilon	29155.4	7.8	1.7	22.1	Cytoplasm, Melanosome	Acetylation, Phosphoprotein	2.7	66.3
P19338	NUCL_HUMAN Nucleolin	76568.5	7.6	1.7	22.2	Nucleolus, Cytoplasm	Acetylation, Methylation, Phosphoprotein	9.4	21.1
Q06830	PRDX1_HUMAN Peroxiredoxin-1	22096.3	10.3	2.4	22.9	Cytoplasm, Melanosome	Acetylation, Disulfide bond, Phosphoprotein	12.3	17.3
Q15084	PDIA6_HUMAN Protein disulfide-isomerase A6 precursor	48091.3	9.4	2.2	23.0	ER lumen, Cell membrane, Melanosome	Disulfide bond, Phosphoprotein	9.5	0.9

P53618	COPB_HUMAN Coatomer subunit beta	107073.8	4.4	1.0	23.1	Cytoplasm, Golgi apparatus membrane, Peripheral membrane, COPI-coated vesicle membrane, ER-Golgi intermediate compartment	Acetylation	6.3	43.8
O75874	IDHC_HUMAN Isocitrate dehydrogenase [NADP] cytoplasmic	46629.6	5.4	1.2	23.1	Cytoplasm, Peroxisome	Acetylation	5.1	9.4
Q16836	HCDH_HUMAN Hydroxyacyl-coenzyme A dehydrogenase, mitochondrial	34255.9	4.9	1.1	23.2	Mitochondrion matrix	Acetylation	8.8	74.0
Q86VP6	CAND1_HUMAN Cullin-associated NEDD8-dissociated protein 1	136288.8	8.1	1.9	23.2	Cytoplasm, Nucleus	Acetylation, Phosphoprotein	3.9	54.2
P08195	4F2_HUMAN 4F2 cell-surface antigen heavy chain	57909.0	9.8	2.3	23.3	Apical cell membrane, Melanosome	Acetylation, Disulfide bond, Glycoprotein, Isopeptide bond, Phosphoprotein, Ubl conjugation	6.2	36.5
P07108	ACBP_HUMAN Acyl-CoA-binding protein	10038.0	4.7	1.1	23.3	ER, Golgi apparatus	Acetylation, Phosphoprotein	2.7	45.2

Q96FW1	OTUB1_HUMAN Ubiquitin thioesterase OTUB1	31264.4	5.1	1.2	23.5	Cytoplasm	Acetylation, Phosphoprotein	2.0	60.6
P13010	KU86_HUMAN ATP-dependent DNA helicase 2 subunit 2	82652.4	8.0	1.9	23.7	Nucleolus, Chromosome	Acetylation, Phosphoprotein, Ubl conjugation	5.7	27.9
P07339	CATD_HUMAN Cathepsin D precursor	44523.7	6.2	1.5	24.0	Lysosome, Melanosome, Secreted	Disulfide bond, Glycoprotein, Zymogen	4.3	32.5
P07900	HS90A_HUMAN Heat shock protein HSP 90-alpha	84606.7	35.8	8.6	24.0	Cytoplasm, Melanosome	Acetylation, Nitration, Phosphoprotein, S-nitrosylation, Ubl conjugation	35.5	3.6
P04075	ALDOA_HUMAN Fructose-bisphosphate aldolase A	39395.3	13.4	3.2	24.2		Acetylation, Phosphoprotein	32.1	148.2
Q14204	DYHC1_HUMAN Dynein heavy chain 1, cytoplasmic 1	532071.8	17.5	4.2	24.3	Cytoskeleton	Acetylation, Phosphoprotein	10.3	42.4
Q99497	PARK7_HUMAN Protein DJ-1	19878.5	9.4	2.3	24.3	Cytoplasm, Nucleus, Mitochondrion	Isopeptide bond, Oxidation, Phosphoprotein, Ubl conjugation, Zymogen	6.3	32.6
P26641	EF1G_HUMAN Elongation factor 1-gamma	50087.2	6.7	1.6	24.5		Acetylation	4.3	31.6

P31939	PUR9_HUMAN Bifunctional purine biosynthesis protein PURH	64575.5	8.2	2.0	24.6		Acetylation	2.2	71.8
P23526	SAHH_HUMAN Adenosylhomocysteinase	47685.3	7.0	1.7	24.7	Cytoplasm, Melanosome	Acetylation		
P53396	ACLY_HUMAN ATP-citrate synthase	120762.1	9.0	2.2	24.8	Cytoplasm, Melanosome	Acetylation	12.4	38.1
P11142	HSP7C_HUMAN Heat shock cognate 71 kDa protein	70854.4	38.3	9.6	25.1	Cytoplasm, Melanosome, Nucleolus	Acetylation, Methylation, Phosphoprotein, Ubl conjugation	24.6	32.5
P38646	GRP75_HUMAN Stress-70 protein, mitochondrial precursor	73634.8	22.3	5.6	25.3	Mitochondrion, Nucleolus	Acetylation	20.1	10.6
P30041	PRDX6_HUMAN Peroxiredoxin-6	25019.2	11.7	3.2	27.5	Cytoplasm, Lysosome, Cytoplasmic vesicle	Acetylation, Disulfide bond, Phosphoprotein	7.2	38.6
P32119	PRDX2_HUMAN Peroxiredoxin-2	21878.2	5.8	1.6	27.8	Cytoplasm	Acetylation, Disulfide bond	3.7	38.2
	Sum		**993.2**	**60.3**	**6.1**				

Table 4C. Actin, Tubulin, GAPDH (G3P) in the nuclear fractions.

Swissprot	Protein name	MW	Avg	SD	CV	Localization	PTMs	Avg SKBR3	% Change
P62736	ACTA_HUMAN Actin, aortic smooth muscle	41981.8	19.6	5.0	25.4	Cytoskeleton	Acetylation, Methylation, Oxidation	7.3	63.1
P60709	ACTB_HUMAN Actin, cytoplasmic 1	41709.7	47.1	17.9	38.0	Cytoskeleton	Acetylation, Methylation, Oxidation, Ubl conjugation	22	52.8
	Sum actin		**66.6**	**21.2**	**31.9**				
Q13885	TBB2A_HUMAN Tubulin beta-2A chain	49875.0	6.8	1.4	21.4	Cytoskeleton	Phosphoprotein	11.4	68.7
Q71U36	TBA1A_HUMAN Tubulin alpha-1A chain	50103.7	8.7	2.3	26.4	Cytoskeleton	Acetylation, Nitration, Phosphoprotein	13	49.5
	Sum tubulin		**15.5**	**3.2**	**32.5**				
P04406	G3P_HUMAN Glyceraldehyde-3-phosphate dehydrogenase	36030.4	9.8	3.2	32.5	Cytosol, Perinuclear region, Membrane	ADP-ribosylation, Acetylation, Methylation, Nitration, Phosphoprotein S-nitrosylation, Ubl conjugation	75	673

Table 4D. Actin, Tubulin, GAPDH (G3P) in the cytoplasmic fractions.

Swissprot	Protein name	MW	Avg	SD	CV	Localization	PTMs	Avg SKBR3	% Change
P62736	ACTA_HUMAN Actin, aortic smooth muscle	41981.8	16.1	2.9	17.8	Cytoskeleton	Acetylation, Methylation, Oxidation	7.8	50.7
P60709	ACTB_HUMAN Actin, cytoplasmic 1	41709.7	22.6	6.2	27.3	Cytoskeleton	Acetylation, Methylation, Oxidation, Ubl conjugation	21.2	6.3
	Sum actin		**38.7**	**7.1**	**18.4**				
P07437	TBB5_HUMAN Tubulin beta chain	49639.0	13.4	1.7	12.9	Cytoskeleton	Acetylation, Isopeptide bond, Methylation, Phosphoprotein, Ubl conjugation	14	4.3
P68371	TBB2C_HUMAN Tubulin beta-2C chain	49799.0	16.5	2.5	14.9	Cytoskeleton	Acetylation, Phosphoprotein	12.8	22.9
Q9H4B7	TBB1_HUMAN Tubulin beta-1 chain	50294.6	9.3	1.7	18.5	Cytoskeleton	Phosphoprotein	6.1	35.9
Q13509	TBB3_HUMAN Tubulin beta-3 chain	50400.3	8.3	1.7	20.8	Cytoskeleton	Phosphoprotein	-	-
Q13885	TBB2A_HUMAN Tubulin beta-2A chain	49875.0	29.0	6.1	20.9	Cytoskeleton	Phosphoprotein	22.2	23.8

Q71U36	TBA1A_HUMAN Tubulin alpha-1A chain	50103.7	28.9	5.3	18.2	Cytoskeleton	Acetylation, Nitration, Phosphoprotein	24.4	15.5
	Sum tubulin		**76.5**	**10.5**	**13.7**				
P04406	G3P_HUMAN Glyceraldehyde-3-phosphate dehydrogenase	36030.4	56.0	8.0	14.3	Cytosol, Nucleus, Perinuclear region, Membrane, Cytoskeleton	ADP-ribosylation, Acetylation, Methylation, Nitration, Phosphoprotein, S-nitrosylation, Ubl conjugation	194	241.2

Table 5. Spectral counts of bovine protein spikes used for assessing experimental variability.

SwissProt ID	Protein name	MW	Avg counts	SD	CV		Avg counts	SD	CV
			MCF7/MCF10 cytoplasmic				MCF7/MCF10 nuclear		
P02666	CASB_BOVIN Beta-casein	25091.3	5.6	0.3	6.1		7.1	1.4	19.6
P02662	CASA1_BOVIN Alpha-S1-casein	24513.4	14.5	1.3	9.0		15.0	1.6	10.7
P01966	HBA_BOVIN Hemoglobin alpha	15043.9	15.3	2.1	13.9		17.5	3.1	17.8
P02070	HBB_BOVIN Hemoglobin beta	15944.3	13.7	2.4	17.5		19.1	7.9	41.4
P02663	CASA2_BOVIN Alpha-S2-casein	26002.3	3.7	1.0	27.9		5.0	1.6	31.7
P00711	LALBA_BOVIN Alpha-lactalbumin	16235.9	4.0	1.6	40.5		7.7	2.1	27.1

4.5 References.

1. Hung, M.C. and W. Link, *Protein localization in disease and therapy.* J Cell Sci, 2011. **124**(Pt 20): p. 3381-92.

2. Lee, K., Byun, K., Hong, W., Chuang, H. Y., Pack, C. G., Bayarsaikhan, E., Paek, S. H., Kim, H., Shin, H. Y., Ideker, T., Lee, B. *Proteome-wide discovery of mislocated proteins in cancer*. Genome Res, 2013.

3. Bossi, A. and B. Lehner, *Tissue specificity and the human protein interaction network.* Mol Syst Biol, 2009. **5**: p. 260.

4. Safran, M., Solomon, I., Shmueli, O., Lapidot, M., Shen-Orr, S., Adato, A., Ben-Dor, U., Esterman, N., Rosen, N., Peter, I., Olender, T., Chalifa-Caspi, V., Lancet, D. *GeneCards 2002: towards a complete, object-oriented, human gene compendium.* Bioinformatics, 2002. **18**(11): p. 1542-3.

5. Mazzola, J.L. and M.A. Sirover, *Subcellular localization of human glyceraldehyde-3-phosphate dehydrogenase is independent of its glycolytic function.* Biochim Biophys Acta, 2003. **1622**(1): p. 50-6.

6. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., Sullivan, M. *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.* Nucleic Acids Res, 2012. **40**(Database issue): p. D261-70.

7. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.

8. Usaite, R., Wohlschlegel, J., Venable, J. D., Park, S. K., Nielsen, J., Olsson, L., Yates Iii, J. R. *Characterization of global yeast quantitative proteome data generated from the*

wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods. J Proteome Res, 2008. **7**(1): p. 266-75.

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

10. Kuster, B., Schirle, M., Mallick, P., Aebersold, R. *Scoring proteomes with proteotypic peptide probes.* Nat Rev Mol Cell Biol, 2005. **6**(7): p. 577-83.

11. McMahon, L. W., Sangerman, J., Goodman, S. R., Kumaresan, K., Lambert, M. W. *Human alpha spectrin II and the FANCA, FANCC, and FANCG proteins bind to DNA containing psoralen interstrand cross-links.* Biochemistry, 2001. **40**(24): p. 7025-34.

12. Perez-Munive, C. and S. Moreno Diaz de la Espina, *Nuclear spectrin-like proteins are structural actin-binding proteins in plants.* Biol Cell, 2011. **103**(3): p. 145-57.

13. Prosniak, M., Dierov, J., Okami, K., Tilton, B., Jameson, B., Sawaya, B. E., Gartenhaus, R. B. *A novel candidate oncogene, MCT-1, is involved in cell cycle progression.* Cancer Res, 1998. **58**(19): p. 4233-7.

14. Hsu, H.L., B. Shi, and R.B. Gartenhaus, *The MCT-1 oncogene product impairs cell cycle checkpoint control and transforms human mammary epithelial cells.* Oncogene, 2005. **24**(31): p. 4956-64.

15. Tenga, M.J. and I.M. Lazar, *Proteomic snapshot of breast cancer cell cycle: G1/S transition point.* Proteomics, 2013. **13**(1): p. 48-60.

Chapter 5. Conclusions

In this work, we identified a novel set of 103 housekeeping proteins (34 nuclear and 75 cytoplasmic) for normalization and validation of label-free quantitative mass spectrometry spectral count data. The protein sets exhibited stable expression level in two cell lines (MCF-7 and MCF-10A), in two different cell cycle stages (G1 and S), and in two cellular subfractions (nuclear and cytoplasmic). The cellular location and biological function of these proteins conferred housekeeping roles to this protein set, supporting their selection for normalization purposes. Preliminary experiments aimed at validating the cytoplasmic protein set in SKBR3 cells demonstrated that ~80 % of the proposed proteins preserved the counts within the limits necessary for the selection of differentially expressed proteins at a two-fold change threshold in spectral count values. Among the commonly used proteins for normalization (actin, tubulin and GAPDH), tubulin represented the most stable and reproducible levels, followed by actin. GAPDH, which is frequently used as a loading control for western blot experiments, presented a stable level of spectral counts in the originally proposed set, but its use for normalization purposes could not be validated due to excessive variability in follow-up experiments with SKBR3 cells. The proposed protein set presents particular value to the development of improved label-free quantitative proteomic data processing methods that will advance the analysis of complex biological samples in which conventional endogenous proteins do not behave in a satisfactory manner. In addition to their applicability for data normalization or validation, a number of members of the protein set present potential utility as cellular markers.