Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Bioinformatics methods and applications for functional analysis of mass spectrometry based proteomics data

Chanchal Kumar

aus

Masimpur-Assam, India

2008

Erklärung

Diese Dissertation wurde im Sinne von §13 Abs. 3 der Promotionsordnung vom 29. Januar 1998 von Herrn Prof. Matthias Mann betreut.

Ehrenwörtliche Versicherung

Diese Dissertation wurde selbständig, ohne unerlaubte Hilfe erarbeitet.

München, am 16.10.2008

_____

(Chanchal Kumar)

Dissertation eingereicht am 16.10.2008

1. Gutachter: Prof. Dr. Matthias Mann

2. Gutachter: Prof. Dr. Karsten Suhre

Mündliche Prüfung am 14.11.2008

"From the standpoint of daily life, however, there is one thing we do know: that we are here for the sake of each other - above all for those upon whose smile and well-being our own happiness depends, and also for the countless unknown souls with whose fate we are connected by a bond of sympathy. Many times a day I realize how much my own outer and inner life is built upon the labors of my fellow men, both living and dead, and how earnestly I must exert myself in order to give in return as much as I have received."

-Albert Einstein

Table of Contents

Summary

With the dawn of the 'omics' era, bioinformatics has been catapulted from being a passive service component of molecular biology to a multifaceted scientific discipline that now actively drives major biomedical endeavors. In the past decade bioinformatics has played key roles in the success of genomics and seamlessly integrated itself into the fabric of contemporary biology. Owing to recent advances in mass spectrometry (MS) instrumentation, proteomics too has joined in the league of high throughput technologies[1,2]. Modern proteomics experiments generate massive amount of data of complex structure and high dimensionality. Analysis of such datasets presents many novel challenges hitherto unknown to proteomics researchers, therefore bioinformatics is gaining wider acceptance in proteomics research. This thesis applies bioinformatics to systematic knowledge mining and comprehensive functional analysis of mass spectrometry based proteomics datasets.

Proteomics in itself builds upon an arsenal of innovative analytical, technological and molecular biology methodologies. It is important to appreciate how these diverse technologies and methodologies coexist and co-operate to facilitate high throughput investigations at the protein level. Breakthroughs in mass spectrometric instrumentation and protein ionization techniques have played pivotal roles in the advancement of proteomics, which has been timely supported by innovations in experimental strategies. With the introduction of various quantitation methods, mass spectrometry based proteomics is now ready for systems-wide measurement of cellular protein expression levels[2]. Chapters 1 through 3 of the thesis provide a brief introduction to these aspects of proteomics, namely mass spectrometry instrumentation, ionization techniques and quantitative proteomics techniques.

The data structures generated by mass spectrometers are a collection of mass spectra, which form a three dimensional space of mass-to-charge ratio (m/z), time (t) and intensity (I). In simpler terms these mass spectra are the signal generated by the ionized peptides from digested proteins and contain peptide identification and quantitation information. The process of decoding peptide identity and quantity from collection of mass spectra is an intensive multi-level algorithmic

exercise, now widely studied in the sub-discipline of *Computational Proteomics*. Computational proteomics spans a gamut of computational, statistical and machine learning methods and algorithms especially dedicated for peptide (and protein) identification and quantitation. Moreover, it harbors a few of the most exciting algorithmic research problems in biology for data disambiguation, interpretation and presentation, and is essentially the cornerstone of proteome informatics[3]. Chapter 4 of the thesis briefly discusses this very important facet of the proteomics workflow.

The analysis of raw mass spectrums by computational proteomics algorithms and applications usually generate a data matrix containing protein identity and quantity information for different biological conditions or samples. Essentially, this inventory of identified proteins and their quantitative map (if present) contain a wealth of information that needs to be mapped onto biological knowledge and insights. This transformation from the data to the knowledge domain is facilitated by bioinformatics. Bioinformatics itself has come to prominence in the last decade due to the large scale genome sequencing projects. But application of bioinformatics to solve biological problems is not a recent trend and can be traced back to the earliest days of the computing revolution[4,5]. As this thesis primarily centers on bioinformatics application, an understanding of current state-of-the-art in this field will help put proteomics related bioinformatics activities in broader context of integrative systems biology. Chapter 5 discusses some of the major research directions in bioinformatics which is now also becoming an integral part of the regular proteomics workflow.

This thesis discusses four projects illustrating wide-ranging functional analysis of mass spectrometry based proteomics datasets by bioinformatics. The data generated by current proteomics research endeavors are mainly of two types - qualitative whereby only the qualitative aspect of the constituent proteome is studied, and quantitative where in addition to cataloguing proteins the quantitative information of their changes across conditions are also reported. The scope of functional analysis and the analytical directions that can be taken are largely dependent on the type of data generated. In qualitative proteomics most of the bioinformatics activities are focused on functional data mining by integration of various annotational data sources - to extract the global biological theme underlying the proteome. Additionally, in quantitative proteomics

machine and statistical learning approaches can be employed to explore quantitative dimensions of the proteome datasets that are not initially obvious to humans.

The first project of this thesis showcases the breadth of biological knowledge that can be extracted from a proteome inventory by applying an array of bioinformatics tools and algorithms. In this project the proteome of the 3T3-L1 adipocyte (a fat cell line) was studied in-depth after fractionation into four sub-cellular fractions, namely cytosol, nuclei, mitochondria and membrane. State-of-the-art MS-based protein identification technology developed recently in our laboratory allowed us to identify more than 3,200 proteins with essentially no false positives. Extensive bioinformatics analysis and comparison with transcriptome data revealed several layers of information related to the adipocytes proteome that were in turn mapped to an ensemble of interesting biological processes, functions and pathways. Our findings concur with recent scientific re-evaluation of adipocytes function and pathophysiology, which renounces the view that they are mere lipid depots and implicates them in myriad cellular and organismal processes[6,7]. Additionally, by using a systemic protein prioritization methodology described recently for functional genomics[8] we predicted candidate proteins hitherto not known to be involved in insulin-dependent vesicular trafficking. Chapter 6 discusses the proteomics and bioinformatics analysis of the 3T3-L1 adipocyte along with key findings in detail.

The second project relates to systems level quantitative proteomic comparison of the mouse hepatoma cell line Hepa1-6 with the non-transformed mouse primary hepatocytes. The experiment was performed by employing the Stable Isotope Labeling by Amino Acids in Cell culture (SILAC) approach[9], whereby Hepa1-6 was completely labeled by the 'heavy' $^{13}C_6$-forms of arginine and lysine and combined with the primary hepatocytes. To characterize the features of these two proteomes, quantitation information (i.e. protein ratios between the two cell types) was used to divide all proteins into five quantiles. Each quantile was clustered according to the Gene Ontology (GO) and KEGG pathway database information to assess their enriched functional groups and signaling pathways. To integrate this information at the systems level, hierarchical clustering based on the enrichment *p*-value obtained from GO and KEGG clustering was performed. Using this novel bioinformatics algorithm for functional data mining, the proteomic phenotypes of the primary cells and transformed cells are immediately apparent.

Primary hepatocytes are enriched in mitochondrial functions such as metabolic regulation and detoxification, as well as liver functions with tissue context such as secretion of plasma and low-density lipoprotein (LDL). In contrast, the transformed cancer cell line Hepa1-6 is enriched in cell cycle and growth related functions. Interestingly, several aspects of the molecular basis of the "Warburg effect" described in many cancer cells became apparent in Hepa1-6, such as increased expression of glycolysis markers and decreased expression of markers for tricarboxylic acid (TCA) cycle[10]. Many of these bioinformatics findings could directly be verified at cellular compartment level by light microscopic comparison of the cell types. This is the first systematic proteome level evaluation of a cell line model against its cognate *in vivo* cell counterpart and chapter 7 discusses the experimental procedure, bioinformatics steps and results obtained.

The third project banks on bioinformatics analysis of one of the most important aspects of any biological process, which is time course progression through a cellular event. In this project the SILAC approach was combined with latest MS technology, large scale phosphopeptide enrichment, sophisticated computational proteomics and novel bioinformatics - to study time course progression of cell cycle in HeLa cell line at the proteome and phosphoproteome level. The dataset comprised of 6 time points capturing different temporal stages of cell cycle and was analyzed separately by a novel time course guided supervised clustering approach. The clusters recapitulated known behavior of key cell cycle players but also revealed interesting patterns of biological functions when analyzed by an extended version of the proteomic phenotypic method described in chapter 7. Moreover, we used directional statistics methods to reveal the cellular component level organization of the proteome during the course of cell cycle progression. Comparison of proteome changes with transcriptome showed interesting regulatory aspects wherein proteins regulated at both the mRNA and the protein level showed an enrichment of cell cycle related functions, whereas proteins regulated at neither of the levels were preferentially involved in homeostatic and basic metabolic processes. Additionally, analysis of phosphoproteome showed time dependent regulation of cell cycle regulated kinase substrates, and in particular identified MCM6 protein as a substrate for the DNA damage response network. More interestingly, systems level integration of proteome, transcriptome and phosphoproteome revealed that E2F transcriptional targets are regulated during the cell cycle through lamina

association. Chapter 8 discusses the experimental protocols, bioinformatics analysis and results obtained from this global cell cycle phosphoproteomics study.

The fourth project explores the spatial aspect of cellular proteomes to elucidate the mitochondrial proteome of mouse brown adipose tissue (BAT) and white adipose tissue (WAT). Here we build on the strength of quantitative proteomics by SILAC and complement it by a novel bioinformatics approach for sub-cellular localization prediction. In a set of two parallel experiments we mix mitochondrial fraction from each of the adipose tissue subtypes (brown and white) with either SILAC labeled nuclear or SILAC labeled post mitochondrial fraction from the cognate cell types (brown adipocyte, 3T3-L1 adipocyte). The resultant relative quantitative ratios were then modeled as bimodal Gaussian distributions, and subsequently used to assign localization probability to the proteins based on their quantitative ratios using Bayesian framework. This enabled us to disentangle the mitochondrial proteome from nuclear or post mitochondrial populations. Our prediction results concurred very well with known Gene Ontology mitochondrial annotations – 94% of those proteins were sorted correctly. We then used this compendium of mitochondrial proteins to filter a separate *in vivo* quantitative mitochondrial proteome of BAT versus WAT. This quantitative proteomic map of adipose tissue mitochondria provided interesting insights into the divergence of key metabolic processes and pathways. For instance, strongly up-regulated pathways in BAT mitochondria were ubiquinone biosynthesis, oxidative phosphorylation, and citrate cycle. In WAT mitochondria systemic up-regulation of pathways was less pronounced, and significantly up-regulated pathways were androgen/estrogen metabolism, fatty acid synthesis, pyruvate metabolism, and – interestingly - metabolism of xenobiotics. Chapter 9 discusses the experimental protocols, bioinformatics analysis and results obtained from this mitochondrial organellar proteomics study. In addition to the work discussed here, I contributed bioinformatics expertise and analysis to several projects in the Department of Proteomics and Signal Transduction at the Max Planck Institute for Biochemistry. The list of publication generated in my thesis's work so far can be found at the end of this summary.

The studies reported in this thesis exemplify extensive applications of bioinformatics tools and algorithms for analysis of high throughput proteomics data. The work presented here samples

important directions in proteomics related bioinformatics research. There are other analytical directions and applications that are currently being reported in the literature and they have been cited at appropriate places throughout the thesis. Extensive application of bioinformatics to proteomics is a relatively recent development, and has been particularly driven by the very large amount of data that is being generated by current proteomics studies. This on one hand has opened newer vistas for data analysis and knowledge mining, and on another presented novel challenges to biomedical informatics researchers. It is rapidly becoming apparent that bioinformatics is indispensable part of the contemporary high throughput proteomics workflows, and that it will continue to play an integral role in proteomics research. We envisage that the role of bioinformatics in proteomics will evolve where not only it will be at the core of functional analysis, but will also provide important pointers for hypothesis generation and testing. Concurrent to developments in bioinformatics, proteomics will also advance to generate fine grained and comprehensive data, amenable for integrative systems level analysis and exploration. Proteomics equipped with bioinformatics is poised to change the outlook of systems biology and is now ready to engender profound changes in biomedical and translational research. Chapter 10 discusses some of the current challenges for bioinformatics research especially in context of its applications to proteomics, and provides perspectives for its symbiosis with mass spectrometry based proteomics in future.

Publications (§ Equal first author contribution)

1. Forner F, **Kumar C**, Luber CA, Klingenspor M, Mann M
   Pathway analysis of mitochondria in brown versus white adipocytes by quantitative proteomics.
   (Manuscript under submission)

2. Olsen JV[§], Vermeulen M[§], Santamaria A[§], **Kumar C**[§], Miller ML, Jensen LJ, Gnad F, Cox J, Jensen TS, Nigg EA, Brunak S, Mann M
   A systems view of the cell cycle by quantitative phosphoproteomics.
   (Manuscript under submission)

3. Pan C[§], **Kumar C**[§], Bohl S, Klingmüller U, Mann M
   Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type specific functions.
   (Manuscript accepted for publication in *Mol Cell Proteomics*)

4. Bonaldi T, Straub T, Cox J, **Kumar C**, Beker PB, Mann M
Combined Use of RNAi and Quantitative Proteomics to Study Gene Function in Drosophila.
Mol Cell. 2008 Sept ; 31(5):762-772.

5. Graumann J, Hubner NC, Kim JB, Ko K, Moser M, **Kumar C**, Cox J, Schoeler H, Mann M
SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins.
Mol Cell Proteomics. 2008 Apr; 7(4): 672-683

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

7. Zougman A, Pilch B, Podtelejnikov A, Kiehntopf M, Schnabel C, **Kumar C**, Mann M
Integrated Analysis of the Cerebrospinal Fluid Peptidome and Proteome.
J Proteome Res. 2008 Jan 4; 7(1):386-399.

8. Shi R, **Kumar C**, Zougman A, Zhang Y, Podtelejnikov A, Cox J, Wisniewski JR, Mann M
Analysis of the Mouse Liver Proteome Using Advanced Mass Spectrometry.
J Proteome Res. 2007 Aug; 6(8), 2963-72

9. Adachi J[§], **Kumar C**[§], Zhang Y, Mann M
In-depth Analysis of the Adipocyte Proteome by Mass Spectrometry and Bioinformatics.
Mol Cell Proteomics. 2007 Jul; 6(7):1257-73.

10. Macek B, Mijakovic I, Olsen JV, Gnad F, **Kumar C**, Jensen PR, Mann M
The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis.
Mol Cell Proteomics. 2007 Apr;6(4):697-707.

11. Zhang Y, Zhang Y, Adachi J, Olsen JV, Shi R, de Souza G, Pasini E, Foster LJ, Macek B, Zougman A, **Kumar C**, Wisniewski JR, Jun W, Mann M
MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes.
Nucleic Acids Research. 2007 Jan;35(Database issue):D771-9.

12. Olsen JV, Blagoev B, Gnad F, Macek B, **Kumar C**, Mortensen P, Mann M
Global, In vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks.
Cell. 2006, 127(3):635-648

13. Adachi J, **Kumar C**, Zhang Y, Olsen JV, Mann M
The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins.
Genome Biology. 2006, 7:R80

1. Mass spectrometry-based proteomics

The 'central dogma' of molecular biology is that biological information flows from DNA to RNA to protein by well defined molecular processes or 'algorithms', whereas proteins do not change the genetic code. With some notable exceptions, all living cells conform to this rule. DNA, the genetic material, also termed the "book of life" contains all necessary information for RNA (and consequently protein) production, and hence serves as the blueprint for building the entire cellular machinery. Therefore, the beginning of the twenty first century witnessed a remarkable scientific project towards deciphering the genetic constitution of organisms. The sequencing of the complete genome of an organism has been a landmark in the history of biomedical research, which was the result of concerted scientific and technological orchestration between various disciplines of biomedical, physical and material sciences.

The efforts and advances made in large-scale sequencing of a large number of genomes, including the human genome, have generated a wealth of useful information, which is revolutionizing our understanding of the complex molecular biology responsible for different physiological processes. Genomics has undoubtedly furthered the human race's endeavors towards global and holistic understanding of organismal function, evolution and disease pathophysiology. However, detailed study of genomes at DNA and RNA levels have also revealed that these digital pieces of information are just one part of the immense and very intriguing scientific puzzle called "life"[11], and may a times not be sufficient to answer many fundamental questions - like the apparent differences between two very closely related species. For instance, when comparing the chimpanzee genome to the human one, focusing on base pair substitutions, more than 98.5 percent of our DNA sequence is identical between the two species[12]. Since the human genome is not very different from that of a chimpanzee, or even a mouse, there must be something else explaining the complexity of the human body and brain. This explanation may largely lie in the gene products, the proteins, which are the active players in living cells in contrast to the static DNA. Supporting this notion, it turns out that the transcriptome and the proteome (especially in the brain) is significantly different between humans and chimpanzees[13]. A similarly interesting example is the discovery that humans have

fewer proteins-coding genes (~ 20,500 ref[14]) than many plants, again adding to the growing body of evidence that the complexity of human species is largely due to multiple protein forms that result from alternative splicing of RNAs[15] and post translational modifications such as phosphorylation, methylation and glycosylation.

With growing appreciation of the dynamic role of proteins in almost every cellular process and direct implication of their altered behavior or malfunctioning in myriad disease pathologies, the task of mapping the protein repertoire of an organism figures prominently on the agenda of twenty first century biomedical science[16,17]. For this reason as well as many others, one of the major challenges in the post-genomic era is to understand the structure, dynamics and interactions of proteins[18]. This knowledge can be acquired in a large-scale fashion by a relatively new "omics"-discipline; namely proteomics[19]. While the origin of proteomics can be traced back to the 1970-80s, it was named by Marc Wilkins in 1994 when he was searching for an alternative to the phrase "the protein complement of the genome"[20, 21].

Proteomics is now broadly defined by the range of technologies that allow large-scale investigation of genetic and cellular function directly at the protein level. Although the field of proteomics builds on an array of analytical, technical and molecular biology methodologies - including protein microarrays[22], global two hybrid analyses[23], and high throughput protein production and crystallization[24], it has in the past decade been particularly driven by the rapid advances made in instrumental technologies, mainly mass spectrometry (MS)[1,25].

1.1 Generic Workflow of MS-based Proteomics

Mass spectrometry is an analytical technique that characterizes a molecule on the basis of the mass to charge ratio (m/z) of its charged gas-phase particles (ions). MS-based proteomics refers to the application of mass spectrometry to the study of proteins. In MS-based proteomics, the m/z values of peptides or intact proteins are measured, which can provide their added amino acid molecular weight and as well as the weight of potential post-translational modifications. Currently, two complementary strategies are commonly employed for characterizing proteins - bottom-up proteomics and top-down proteomics.

The bottom-up approach is the most pervasive and by far the most successful method in proteomics (Figure 1.1). Proteins of interest are enzymatically cleaved at specific sites to yield short peptides of typically 6-20 amino acid residues. The resulting peptide mixture is analyzed in a mass spectrometer in two stages. In the first (survey scans or MS^1 scans), the masses of the intact peptides are determined; in the second, these peptide ions are fragmented by low energy collision with an inert gas (MS/MS fragmentation) to produce amino acid sequence related information. Taking a protein database as a reference, mass spectra are correlated to amino acid sequences with the aid of computational algorithms. The assigned peptide sequences are then mapped to parent proteins, which ultimately leads to protein identification.

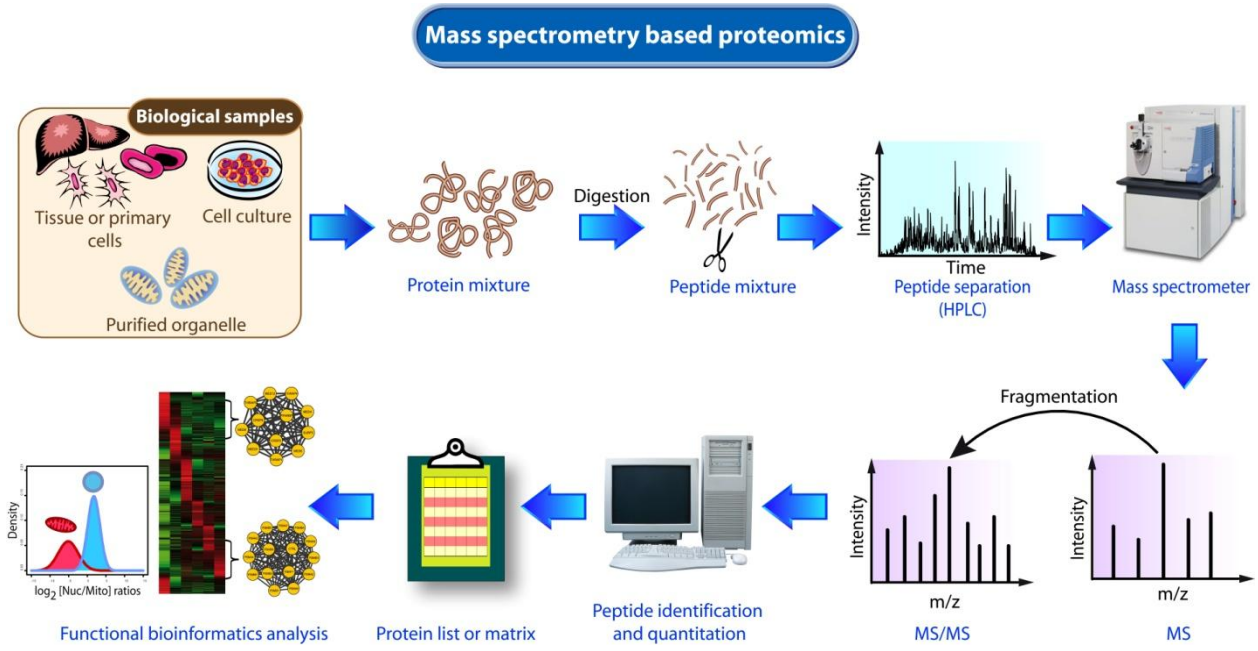


Figure 1.1 General workflow for bottom-up MS-based proteomics. Proteomics samples come from tissues, cell lines, body fluids etc. Protein or peptide samples can be fractionated by different means to reduce complexity and subsequently separated by HPLC. This separated peptide mixture is then introduced into a mass spectrometer for MS and MS/MS analysis thereby generating mass spectra read outs. Computer algorithms match mass spectra to amino acid sequences. The outcome of the experiment is a list of identified proteins which are then used for functional proteomics analysis.

In the top-down approach, intact proteins are ionized and sprayed into mass spectrometers where peptide fragments are subsequently generated using one of a variety of activation methods such

as CID, ECD, ETD, and IRMPD methods (see abbreviation page). Top-down proteomics consists of fragmenting intact proteins in the mass spectrometer, without prior enzymatic digestion. It provides an alternative approach in proteomics to peptide based approaches. It obtains better sequence coverage of protein identifications and is especially suited for elucidating the primary structure, splice variants, and post translational modification (PTM) of whole proteins[26]. So far, the applicability and throughput of this method has been limited, because of its limited sensitivity, the heterogeneity of most proteins and the difficulty of fragmenting proteins efficiently, resulting in prolonged acquisition times and necessitating a combination of different fragmentation techniques. It has also been difficult to combine this methodology with online LC-MS analysis. However, with recent advances in mass spectrometric instrumentation, top-down MS could become a powerful and practical addition to bottom-up in the future[27,28].

Proteomics frequently deals with complex protein or peptide mixtures. As dynamic range and sequencing speed are the limiting factors in current MS technology[29] for dealing with these mixtures. To reduce the complexity of a biological sample (cell lysate, purified organelle, body fluid) it is often desirable to fractionate and separate the proteins or peptides by at least two orthogonal separation techniques before MS analysis. Separation with reversed phase chromatography (HPLC) connected on-line to the mass spectrometer has proven to be a useful method (Figure 1.1).The C_{18} reversed phase HPLC column elutes peptide mixtures with linearly increasing organic solvent, e.g. acetonitrile (MeCN). The gradual elution (typically at a few hundred nanoliter per minute) with shallow gradients increases available sequencing time in the MS. Prior to this hydrophobicity-based peptide separation, complex samples can be separated by one dimensional gel[1], isoelectric focusing[30], ion-exchange[31,32], molecular size[33], and affinity binding such as immunoprecipitation[34,35], IMAC[36,37] and TiO_2 enrichment[29,38,39]. For less complex samples, static electrospray – called nanoelectrospray –can be used which often yields better identification results. In particular static spray is beneficial for targeted modification studies[40].

1.2 Computational and Functional Proteomics

One of the primary goals of any MS-based proteomics experiment is to identify and quantify proteins in a biological sample of interest. This task is performed in the penultimate step of the MS-based proteomics workflow, now widely recognized as "Computational Proteomics"[41]. A large number of computational and algorithmic approaches are now available for identification of proteins and their post translational modifications[42-44] as well as for their quantitation[45]. As further functional analyses rely on the accuracy of protein identification and quantitation, a range of research activities have emerged for accessing the quality of identification and quantitation[46,47].

The ultimate goal of any proteomics endeavor is to gain important insights into the physiology of the biological system under study. Functional proteomics analysis therefore forms the cornerstone of such an enterprise. Data analysis and mining takes place at the end of the work flow and comes under the auspices of "Bioinformatics". Depending upon the nature of the data (qualitative or quantitative) specialized analytical directions are undertaken. Annotation databases such as Gene Ontology[48], pathway resources like KEGG[49], disease databases like OMIM are frequently used to group proteins according to their common biological themes[50-52]. Various statistical and machine learning methods have been employed to analyze proteomics datasets[53-55]. Additionally, integration of other "omics" datasets (genome, transcriptome, epigenome and interactome) can complement the analysis and provide deeper insights into proteome organization, function and dynamics[30,56-60]. The amalgamation of proteomics with other contemporary "omics" disciplines heralds the beginning of a concerted scientific effort by which the future promises of "systems biology" will be realized and delivered[61-63].

2. Mass spectrometry

Mass spectrometry (MS) has become the most important technology in proteomics today[1] because it is a versatile tool that encompasses several unique features at once, such as identification of individual proteins in a complex mixture[51], quantification of proteins in a cell or organism[64] and characterization of important post-translational modifications (PTMs) of proteins (e.g. phosphorylation, methylation, acetylation)[65].

MS is essentially a technique for weighing molecules, but the measurements are obviously not performed with a conventional balance or scale. Instead, the basis of MS is the production of gas-phase ions, which are subsequently separated or filtered according to their mass-to-charge (m/z) ratio in a magnetic or electrostatic field and finally recorded by a detector. The resulting mass spectrum is a plot of the relative abundances of the produced ions as a function of their m/z ratio (Figure 1.1).

Because every peptide molecule and modification has a characteristic mass, MS in proteomics is a very useful and nearly universal tool for analysis, which can provide a nearly unique feature of the molecule. Peptides furthermore have distinct fragmentation patterns that provide structural information to identify their amino acid sequences and modifications.

2.1 Types of ionization and Mass Spectrometers used in proteomics

2.1.1 Electrospray Ionization - ESI (2+, 3+)

In the early 1980s John Fenn and coworkers reported a significant refinement of an ionization principle, originally reported by Malcolm Dole[66] almost two decades earlier. He developed the electrospray ionization (ESI) method as a mass spectrometric technique[67] (Figure 2.1). ESI allows for large, non-volatile molecules (such as peptides and proteins) to be non-destructively ionized directly from a liquid phase (usually a mixture of volatile organic solvent and acidified water). In electrospray, a liquid is passed through a needle to which a high voltage is applied. The charged liquid becomes unstable as it is forced to hold more and more charges. Soon the liquid reaches a critical point and at the tip of the liquid stream in front of the needle it blows apart into a cloud of tiny, highly charged droplets. These droplets rapidly shrink as solvent

molecules evaporate from their surface increasing the electric field at the droplet surface. By a process of 'ion evaporation' (Iribarne and Thomson model) or simple solvent evaporation (charged residue model), the "naked" biomolecule becomes a gas-phase ion.

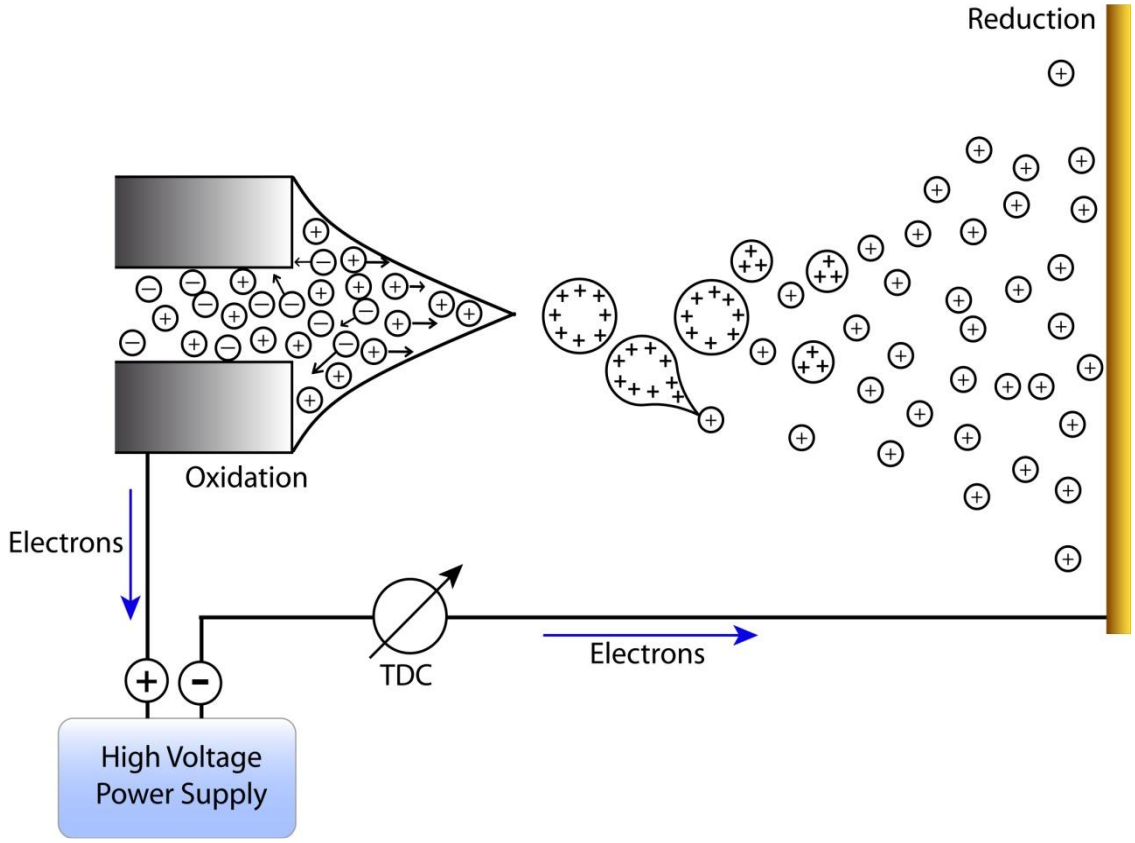


Figure 2.1 Electrospray ionization process schematic (adapted from ref[68])

2.1.2 Matrix-assisted laser desorption ionization –MALDI(1+)

Another "mild" ionization technique for biomolecules is named Matrix-assisted laser desorption ionization (MALDI) and was also developed in the late 1980s (by Franz Hillenkamp and Michael Karas at the University of Münster in Germany)[69](Figure 2.2). In this technique, analyte molecules are co-crystallized with an UV- or IR-absorbing substance termed the matrix, which is usually an organic carboxylic acid such as 2,5-dihydroxybenzoic acid (UV-absorbing) or succinic acid (infrared absorbing). The analytes are desorbed and ionized by a laser beam (pulsed laser irradiation) from the solid surface containing the organic matrix compound in approximately thousand-fold excess. Although the exact ionization process of MALDI, like that

of electrospray, is still not entirely clear, the matrix plays a key role in this technique by absorbing the laser light energy and causing a small part of the target substrate to vaporize. A widely accepted view is that, following their desorption as neutrals, the sample molecules are ionized by acid-base proton transfer reactions with the protonated carboxylic acid matrix ions in a dense 'selvage' phase just above the surface of the matrix.

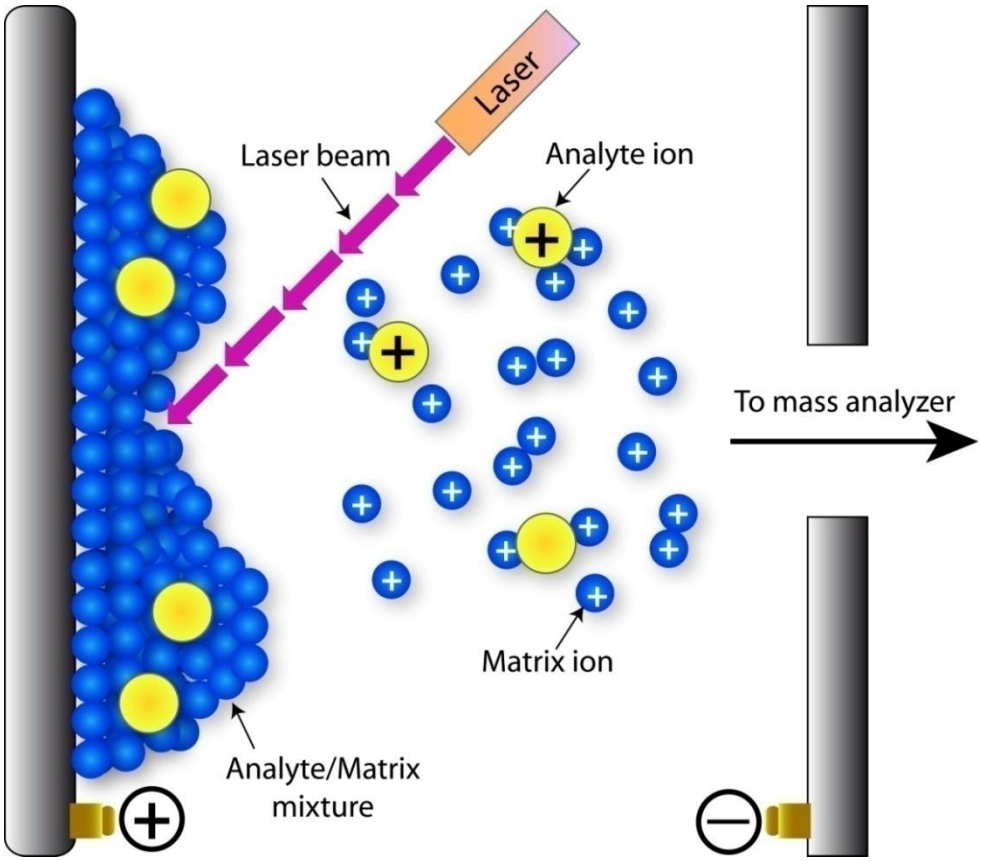


Figure 2.2 MALDI ionization process

The MALDI technique has some similarities with the soft laser desorption (SLD) technique described by Koichi Tanaka in 1987[70]. Tanaka discovered that by mixing ultra fine metal powder in a glycerol matrix an analyte molecule can be ionized resulting in stable gas-phase ions with intact primary structure. Tanaka was awarded the Nobel Prize in Chemistry in 2002 for his work, although this raised a major controversy in the MS community, because most mass spectrometrists felt that the award should have been given to Hillenkamp instead.

Like ESI, MALDI is capable of efficiently ionizing large biomolecules such as peptides and proteins and is often used with time-of-flight (TOF) mass spectrometers due to the vacuum-compatibility and pulsing nature of the technique (the laser frequency can easily be synchronized with the TOF extraction pulse). Both the ESI and MALDI techniques have enabled biological molecules exceeding one million Daltons to be introduced into mass spectrometers[71].

For peptide analysis, the main difference between the two ionization methods is that ESI predominately produces multiply charged ions, MH_n^{n+}, whereas MALDI almost exclusively generates singly-charged peptide ions, MH^+, which can be difficult to sequence by the low-energy dissociation methods available on most proteomic mass analyzers. 2D-gel based proteomics is almost exclusively driven by MALDI-TOF MS analysis, whereas many other areas of proteomics are mainly based on ESI, because of the possibility to couple ESI directly with online LC-MS/MS.

2.2 Traditional mass analyzers in proteomics: TOF, quadrupoles and ion traps

2.2.1 Time-of-flight mass spectrometry

A time-of-flight mass spectrometer (TOF-MS) measures the (mass-to-charge dependent) time it needed for ions of different masses to travel from the ion source region to reach the detector[72]. This requires that the starting time (the time at which the ions leave the ion source) is well defined. Therefore, ions are either formed by a pulsed ionization method such as MALDI or various kinds of rapid electric field switching which are used as a 'gate' to release the ions from the ion source in a very short time interval.

An electric field accelerates all ions into a field-free drift region with a kinetic energy of:

$$E_{kinetic} = q * U$$

Where, q is the ion charge and U is the applied voltage. Since the ion kinetic energy is equal to $E_{kinetic} = \frac{1}{2} * m * v^2$, lighter ions have a higher velocity than heavier ions and reach the detector at the end of the drift region sooner; $q * U = \frac{1}{2} * m * v^2 \Longrightarrow v = \sqrt{\frac{2*q*U}{m}}$

Since the transit time (t) through the drift tube is: $t = \frac{d}{v}$, where d is the length of the drift tube the following equations are used to produce and calibrate a mass spectrum from the signal produce at the detector: $t = \sqrt{\frac{d^2 * m}{2 * q * U}} \Longrightarrow \frac{m}{q} = \frac{2 * U * t^2}{d^2}$

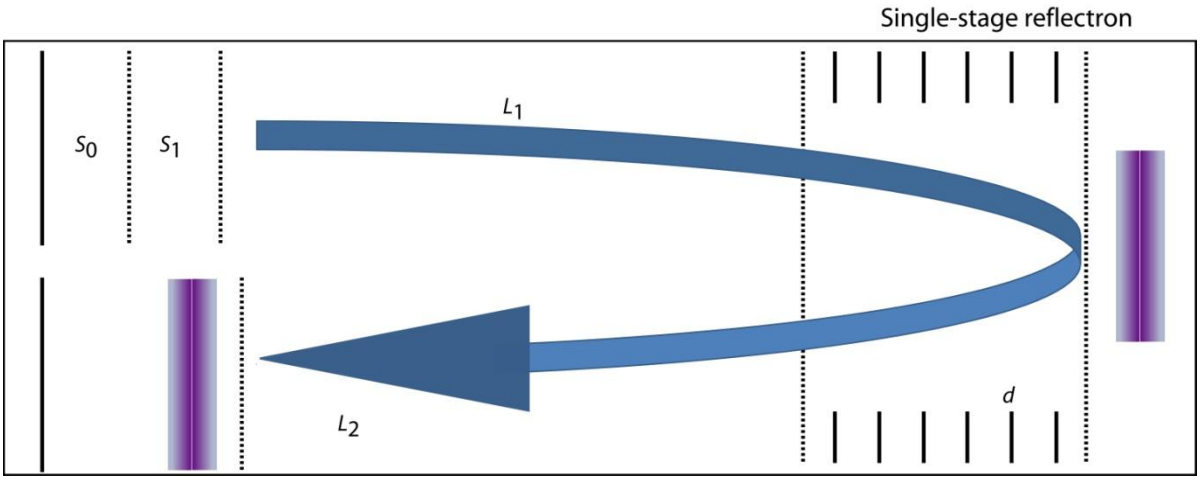


Figure 2.3 Reflectron time-of-flight mass analyser with a dual stage ion extraction source

The ions leaving the ion source of a TOF mass spectrometer have neither exactly the same starting times nor exactly the same kinetic energies, which results in peak broadening and decreased mass accuracy. Various time-of-flight mass spectrometric designs have been developed to compensate for these differences. A reflectron is an ion optic device in which ions in a time-of-flight mass spectrometer pass through a "mirror" or "reflectron" and their flight direction is reversed (Figure 2.3). The reflectron is composed of a series of rings or grids that act as an ion mirror. This mirror compensates for the spread in kinetic energies of the ions as they enter the drift region and improves the resolution of the instrument. The output of an ion detector is measured as a function of time to produce the mass spectrum. Reflector TOF analyzers can easily baseline resolve multiply-charged peptide ions and current instruments used in proteomics usually provide mass accuracy measurements of better than 50 ppm.

2.2.2 Quadrupole ion trap MS

There are two types of ion trap mass analyzers; dynamic and static ion traps. 3-dimensional quadrupole ion traps (QIT) are dynamic traps and Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers contain static traps. Both types of instruments work by trapping ions in electric or magnetic field, respectively, and by manipulating the ions through DC or DC and RF electric fields in a series of carefully timed events. The 3D-ion trap mass spectrometer (also know as the Paul trap, Figure 2.4) was the first integrated and easy to use proteomic analyzer that combined online HPLC with fast scanning tandem MS.

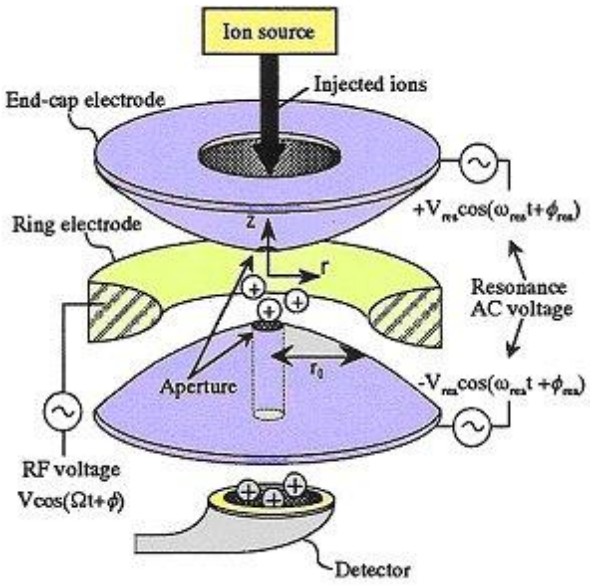


Figure 2.4 Quadrupole ion trap (Paul trap)

The quadrupole ion trap is based on the same principle as a quadrupole mass filter, except that the quadrupole field is generated within a three-dimensional trap. The 3D ion trap consists of a ring electrode and two end caps (Figure 2.5). Ions are formed within the ion trap or (in proteomics) injected into the trap from an external source. The ions are dynamically trapped by the applied RF potentials with the help of a "bath gas" which confines them in the trap. The trapped ions can be manipulated by RF fields analogous to a quadrupole mass spectrometer. A mass spectrum is obtained by changing the electrode voltages to eject the ions from the trap in turn (i.e. scanning the ions out to the detector).

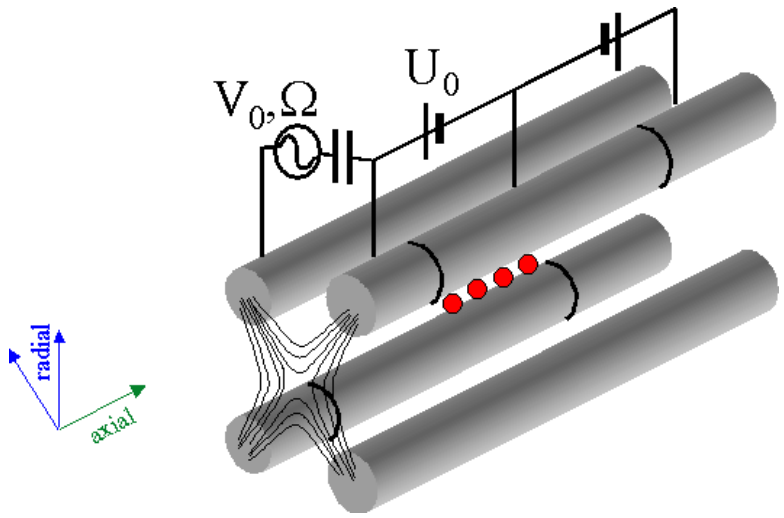


Figure 2.5 Linear ion trap

The Paul trap has very high sensitivity (sub-femtomole) and fast sequencing speed (msec timescale), but large-scale proteomic analysis can be severely hampered by the low mass accuracy and dynamic range of this instrument. The main weakness of Paul traps is their limited dynamic range, which is due to a low trapping and storage capacity (in a point-like three dimensional volume in the middle of the trap) holding less than 10,000 ions before the onset of space-charge distortions.

To overcome these drawbacks of conventional 3D-ion traps, a new generation of ion traps with superior ion capacity, dynamic range, scan speed and sensitive has been introduced[73,74] These are the linear ion traps (or 2D ion traps); essentially segmented rf/dc-quadrupole mass filters, capable of trapping and detecting a factor hundred more ions than traditional 3D-ion traps.

2.3 Hybrid instruments - State-of-the-art MS analyzers

The term hybrid mass spectrometer refers to a combination of two MS analyzer types. In proteomics, the term was originally given to the combination of a quadrupole analyzer with a

time-of-flight detector, namely the qTOF type instruments (Q-TOF and QSTAR[75,76]. The idea is that a hybrid instrument combines the strengths of each analyzer type while minimizing the compromises that might result from interfacing the two or more MS technologies.

In a quadrupole time-of-flight instrument, the high-resolution and sensitivity of the reflector TOF are combined with the tandem MS capabilities of an electrospray quadrupole instrument. This combination represents an instrument that is very well suited for large-scale proteomics projects, and which can easily be coupled to online nano-HPLC for high-throughput analyses. This instrument was a real boost to proteomics when it was introduced in the 1990s. However, today it has been somewhat eclipsed by the latest generation of hybrid mass spectrometers, the linear ion trap – Fourier transform instrument combinations described below.

2.3.1 LTQ-FT - a linear quadrupole ion trap – 7T-FTICR mass spectrometer

Although Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) has shown great promise and potential in proteomics research for more than a decade[77,78,79] it was not until recently, with the introduction of a linear ion trap - Fourier transform mass spectrometer (the LTQ-FT[80]), that FTICR mass spectrometry really became practical and offered adequate sensitivity and speed for large-scale proteomics research (Figure 2.6).

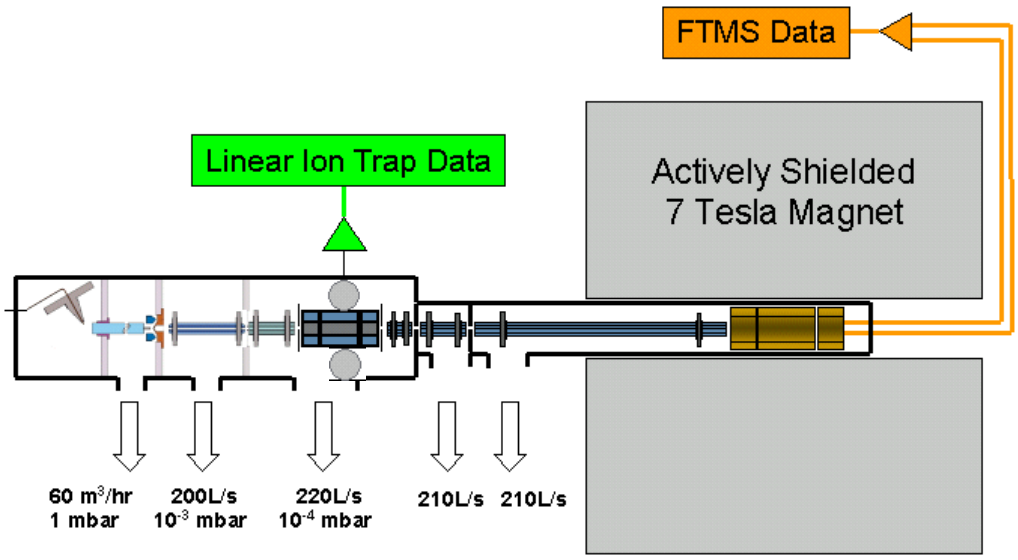


Figure 2.6 High-resolution analyzer: Linear ion trap – FTICR mass spectrometer

The LTQ-FT is a hybrid instrument that can provide sub-ppm mass accuracy with acquisition speed of less than one second per scan. The FTICR-detection (Figure 2.7) combined with automatic-gain control; AGC in the ion trap mass spectrometer allows for controlling the number of ions in the ICR cell during nanoLC-MS measurements, and hence provides high-mass accuracy measurements of peptide ions of less than 2 ppm root-mean-square (RMS) across LC elution profiles. This feature together with its high dynamic range and comparative ease of use makes the hybrid instrument very powerful in identifying peptides via database search algorithms with very high confidence.

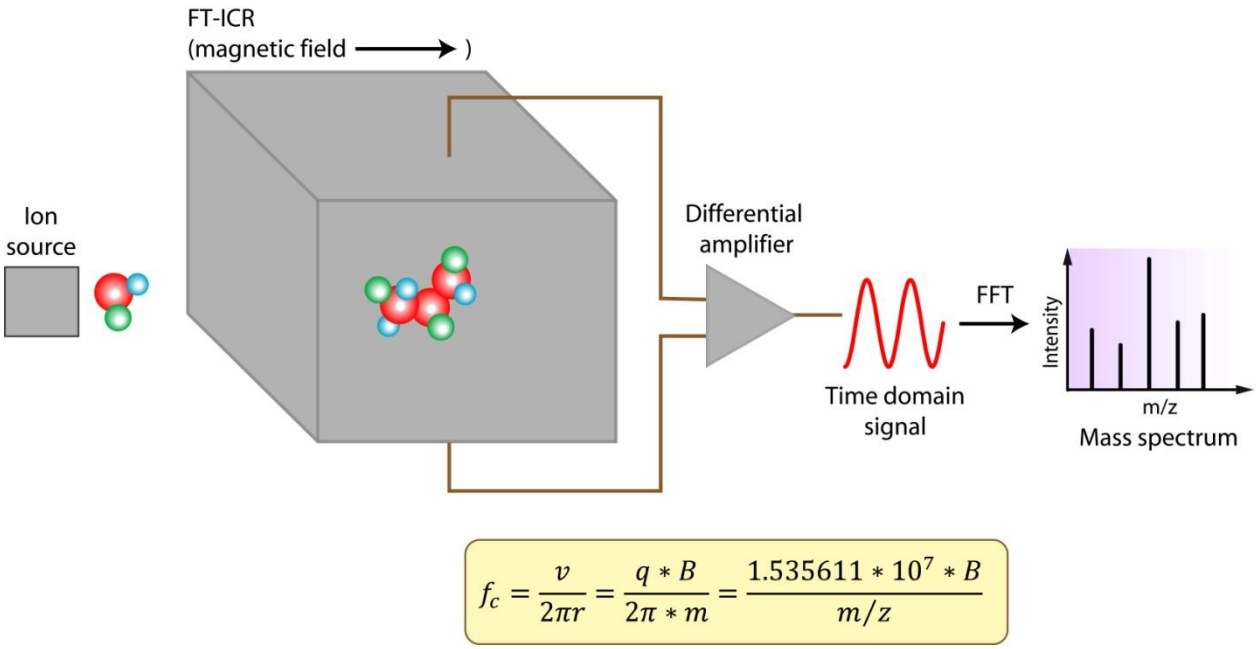


Figure 2.7 Detection principle of an ICR instrument (adapted from ref[81]). The ion cyclotron frequency is directly proportional to the magnetic field and inversely proportional to the mass to charge of the detected ion. Fourier transformation generates a mass spectrum from the overlapping frequency spectrum of several compounds.

In order to maximize the dynamic range and sensitivity of the FT-detector for complex mixture analysis, we and others[82,83] have injected very high numbers of ions into the FT part of the hybrid instrument for FT survey scans. These ion target values are close to the maximum ion capacity of the upstream ion trap of 5-10 E6. However, at high-target values (ACG > 1,000,000) the mass

measurement accuracy of LTQ-FT-MS is greatly affected by space-charge effects in the ICR cell. Space-charging gives rise to a non-linear ion intensity-dependent cyclotron frequency-shift of the ions and as a consequence of the mass measurement accuracy fluctuates 25 ppm or more. This presents a problem for confident peptide identification. To minimize this effect we have developed an acquisition method that utilizes the fast and very sensitive selected ion monitoring (SIM) scan capabilities of the LTQ-FT[84]. Briefly, by accumulating only a small mass range and a limited and defined ion population, space charge effects are eliminated allowing very high mass accuracy and better quantitative accuracy. By employing FT-SIM scans for all parent ions selected for sequencing we have shown that it is possible to measure peptide ions with mass errors less than 2 ppm. On the downside, although the LTQ-FT allows for parallel detection by both ion trap and FT-detectors simultaneously, the SIM scans still takes away valuable analysis time during an online LC-MS/MS experiment.

2.3.2 LTQ-Orbitrap

A major breakthrough in proteomics came very recently with the introduction of a novel mass spectrometer, the LTQ-Orbitrap[85]. The Orbitrap is the first fundamentally new mass analyzer in more than 20 years (Figure 2.8).

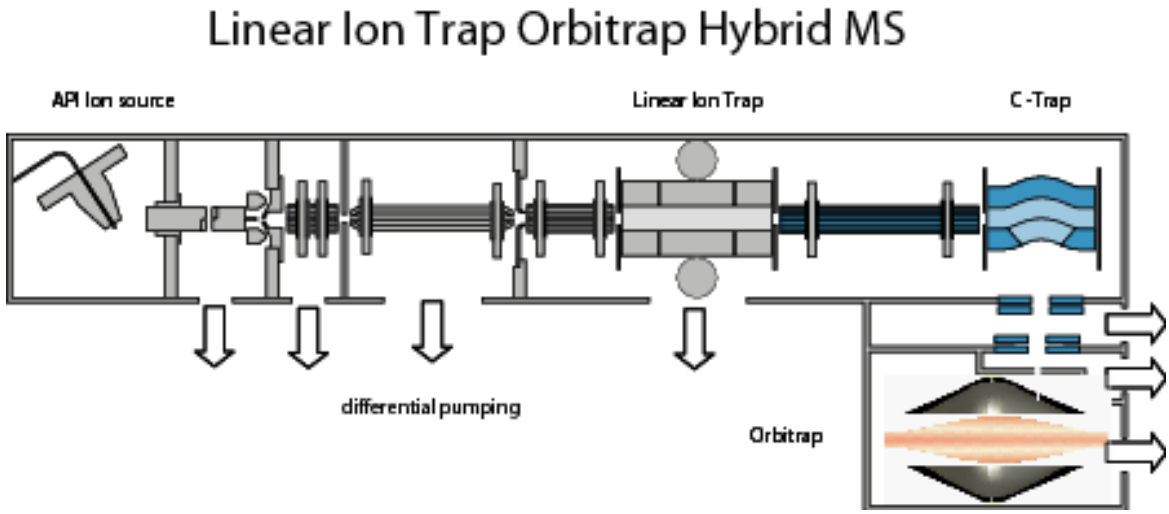


Figure 2.8 LTQ-Orbitrap mass spectrometer

The instrument contains three components. Like the LTQ-FT it has a linear quadrupole ion trap (LTQ), in which it is possible to control and manipulate (e.g. accumulate and collisionally

activate) ions in the sub-second time-scale. Detection can be achieved in two ways. In the linear ion trap ions can be ejected radially through slots (holes) in the quadrupole-rods and detected by two electron multiplier detectors. Alternatively, ions are ejected axially from the ion trap and transferred via octopole-ion guides into another ion trap (the C-trap) were they are collisionally cooled and focused, before they are orthogonally injected into the third component of the instrument, the high-vacuum Orbitrap FT-MS analyzer.

This hybrid instrument exhibits a several-fold increase in sensitivity and dynamic range as compared to a state-of-the-art hybrid ICR mass spectrometer (LTQ-FT). Based on the highly accurate mass determination combined with high resolution and sensitivity, the novel instrument not only allows for routine analysis in a high throughput manner, but also for the straightforward analysis of intact proteins without chemical or enzymatic digestion. Since it is easier to stabilize a magnetic field than an electric field, the LTQ-FT is expected to achieve higher mass accuracies than the LTQ-Orbitrap. However, we developed an internal recalibration method for the LTQ-Orbitrap that offers sub-ppm mass measurement accuracies of peptide ions throughout an online LC-MS/MS experiment even when operating the orbitrap analyzer at high target values[86]. Utilizing a special lock-mass recalibration feature that this hybrid triple-MS instrument provides we are able to recalibrate all ions in FT-MS and MS^n to better than 2 ppm mass error. Briefly, because the C-trap is a second ion trap we can perform multiple ion fillings in this trap and store several "different" ion populations that have been accumulated and axially ejected from the linear ion trap before injecting them into the Orbitrap. The lock-mass recalibration approach takes advantage of this unique feature providing an opportunity to do internal recalibration of all ions by introducing a small number of known recalibration mass ions into every ion population that is filled into and stored in the C-trap. During online nanoLC-MS/MS measurements we chose to make use of a polycyclodimethylsiloxane (PCM) ion as our lock-mass. PCMs are contaminants from ambient air that easily ionize by electrospray ionization and are therefore present throughout the whole chromatographic run. The LTQ-Orbitrap instrument is particularly suitable for both qualitative and quantitative analysis of complex peptide mixtures, because of the high dynamic range and sequencing speed. This allow for sequencing of thousands of peptides by tandem MS in less than one hour of LC-MS/MS analysis. Clearly, the new generation hybrid instruments already have and will continue to set the standards for large-scale

proteomic analyses. Especially the demands for faster sequencing abilities, higher dynamic range, and sensitivity and mass accuracy will be met by these hybrid instruments.

3. Quantitative Proteomics

Any study of biological processes benefits immensely from the knowledge of quantitative change of the entities that form the components of that process. As proteins are the key mediators and final effectors in most of the cellular processes, the study of protein amounts and their quantitative changes has always been an important task in biomedical research. In that context various protein quantitation methods have been developed in the past decades. The emergence of mass spectrometry (MS) as a tool for protein quantitation is a recent development. Protein quantitation by MS depends on measurement of peptide signal intensities and has been mostly done by metabolic incorporation of stable isotopes in proteins or by stable isotope tagging of peptides. Recently comparison of signal intensity of unlabeled peptides by advanced machine learning and statistical models has been developed and is termed label free quantitation[47]. In this chapter we briefly review these methods which are at the core of proteomics workflows.

3.1 Stable isotope dilution

Although mass spectrometry has proven itself as an extraordinary tool for protein characterization, peptides and proteins cannot be directly quantified by MS analysis alone[87]. The reason for this is that peptide responses (signals) in the mass spectrometer are extremely variable, because the ionization efficiency is highly dependent on their chemical structure.

To overcome this problem, proteomic investigators have adapted a quantitative method that has been employed for many years in small-molecule MS, stable isotope dilution[88]. In this procedure heavy stable isotope atom(s), such as deuterium (D), carbon-13 (^{13}C), nitrogen-15 (^{15}N) or oxygen-18 (^{18}O), is covalently incorporated into peptides derived from one of the proteomes to be analyzed by MS. The physical characteristics and properties of the "heavy" stable isotope labeled peptides remain essentially the same as the corresponding "light" versions. When peptides from two cell states are analyzed concurrently (after mixing), they will appear as pairs in the spectrum - spaced by the mass-shift introduced by the isotopic labeling. Relative quantitation is measured by comparing the intensity of the signals of the identical, however isotopically distinct, peptides.

In the last decade several techniques based on stable isotope dilution for protein quantitation by mass spectrometry have emerged. They all enable peptides derived from two or more samples to be distinguished in the mass spectrometer. Many different isotope tag techniques have been described but a few have attracted most of the attention. These include isotope-coded affinity tags (ICAT), metabolic labeling by ^{15}N-incorporation, stable isotope labeling of amino acids in culture (SILAC), enzymatic ^{18}O-labeling and the very recently introduced chemical labeling by tandem mass tags iTRAQ[9, 89-92].

3.2 Isotope coded affinity tags (ICAT)

ICAT™ was the first real breakthrough approach to quantitative expression analysis of complex protein mixtures (Figure 3.1). It was developed in Ruedi Aebersold's laboratory at the University of Washington[89]. ICAT was devised as an LC-MS based approach but it also works with 1D- and 2D-gels.

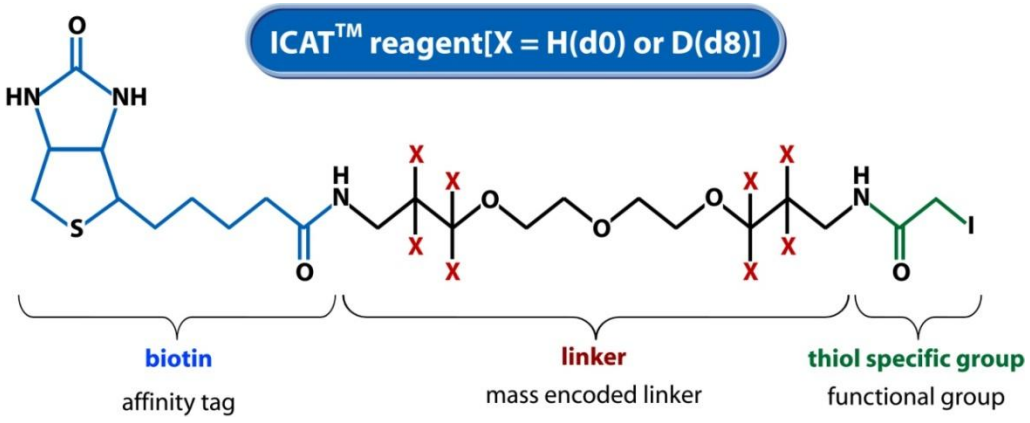


Figure 3.1 Isotope coded affinity tag (ICAT)

The method uses a cysteine thiol-reactive chemical tag that contains biotin, which is used as an affinity tag to specifically isolate the tagged peptides by avidin-affinity chromatography. Although this methodology efficiently reduces the complexity of the protein mixture under investigation, in its original version it had some severe disadvantages. The isotope-mass shift was introduced by the incorporation of eight deuterium atoms in the heavy version of the ICAT-reagent and this resulted in a quantitation problem arising from the fact that the light and heavy labeled version of the same peptide did not co-elute during reversed-phase LC-MS, because of

the so-called deuterium-effect. This RP-LC separation effect is due to the different hydrophobicity of hydrogen (H) and deuterium (D) atoms covalently bound to carbon-atoms (deuterium isotope effects are primarily attributed to the shorter bond length of C-D *versus* C-H, this decreases the hydrophobicity of deuterated peptides). Another problem with the original version of the ICAT-reagent was the generation of very abundant fragment ions from the biotin-group during MS/MS analysis. This made database search identification of ICAT-labeled peptide sequences very difficult. As a consequence of these problems, a new and improved version of the reagent has been introduced with a $^{13}C_9$-label and an internal acid-cleavable bond, which allows removal of the biotin-moiety prior to MS analysis.

3.3 HysTag

Olsen *et al.* developed an isotope coded affinity tag called the HysTag[93], which is an inexpensive and elegant alternative to the ICAT reagent (Figure 3.2). The HysTag reagent is a 10-mer derivatized peptide, H_2N-$(His)_6$-Ala-Arg-Ala-Cys(2-thiopyridyldisulfide)-CO_2H, which consists of four functional elements: an affinity ligand (His_6-tag), a tryptic cleavage site (-Arg-Ala-), an isotope label; Ala-9 residue that contains four (d4) or no (d0) deuterium atoms, as well as a thiol-reactive group (2-thiopyridyl disulfide). For differential analysis cysteine residues in the samples to be compared are modified using either (d4) or (d0) reagent. The HysTag peptide is preserved in Lys-C digestion of proteins and allows charge-based selection of cysteine-containing peptides, whereas subsequent tryptic digestion reduces the labeling group to a di-peptide (-Cys-Ala), which does not hinder effective fragmentation. Surprisingly, it was found that HysTagged peptides containing Ala-d4 co-elute with their d0-labeled counterparts during RP-LC-MS/MS analysis making the HysTag reagent very economical.

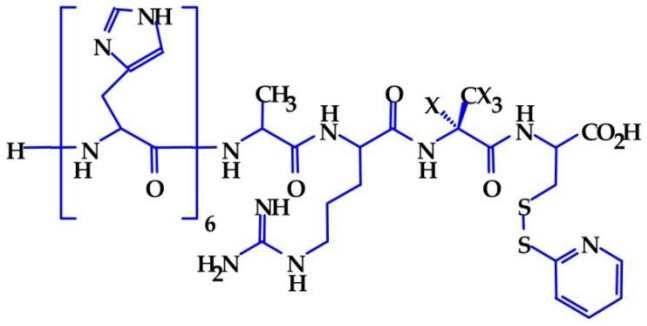


Figure 3.2 The HysTag reagent

The reagent is very easy to manufacture by the well-established solid-phase Fmoc-peptide synthesis and by using a deuterium label the cost of the reagent is low. Olsen *et al.* have successfully studied the plasma membrane proteomes of distinct mouse brain compartments with this approach[93] (Figure 3.3). The HysTag reagent and small scale organellar purification protocols allowed, for the first time, quantification of large numbers of membrane proteins in a single animal.

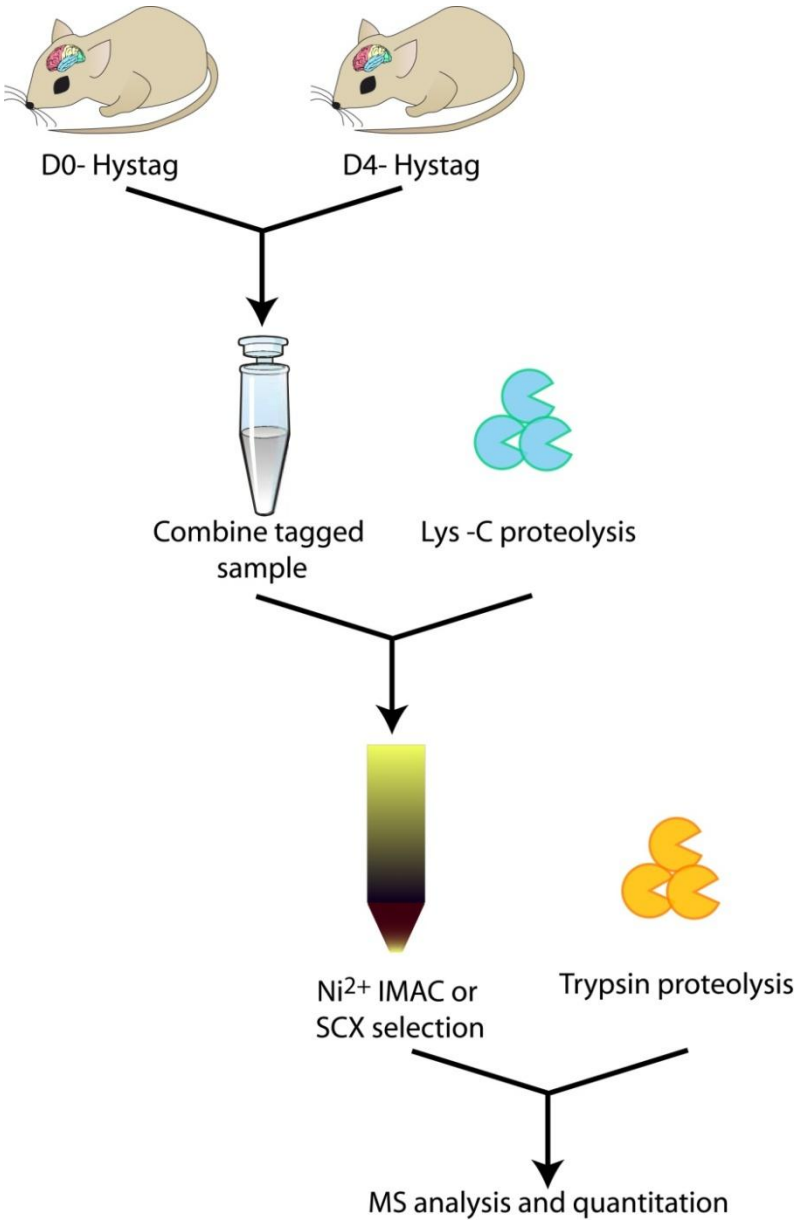


Figure 3.3 The HysTag flowchart for differential analysis of membrane proteins from distinct areas of mouse brain

3.4 Metabolic labeling

There are several different ways to metabolically isotope label cells during growth. Stable isotope incorporation in newly synthesized proteins can be achieved by a stable-isotope labeled compound, which represents the sole source of an element, typically nitrogen (^{15}N) or carbon (^{13}C). This methodology is very simple, but requires a well-defined cell culture system, where the cells of interest are capable of synthesizing all the necessary amino acids from the isotope labeled compound. Hanno Langen introduced this method and applied it to *E. coli* bacteria and yeast cells[94]. Chait and coworkers reported the ^{15}N labeling in yeast for protein quantitation and phosphorylation study[90]. Heck and coworkers reported a two-step approach to ^{15}N-encode higher-multicellular organisms. Metabolic labeling of the model organisms *C. elegans* and *D. Melanogaster* was accomplished by feeding them on completely ^{15}N-labeled *E. coli* or yeast cells[95]. Yates and coworkers even attempted to isotope label an entire mammalian organism, a rat, by feeding it a diet consisting of ^{15}N as the single source of nitrogen[96].

3.5 Stable Isotope Labeling by Amino acids in Cell Culture (SILAC)

Our laboratory described a whole cell metabolic labeling strategy termed stable isotope labeling of amino acids in cell culture ('SILAC')[97]. SILAC makes proteins from one cell population (isotope-encoded with an essential amino acid such as $^{13}C_6$-Arg) distinguishable from proteins from another cell population (cultured in media containing 'normal' $^{12}C_6$-Arg) (Figure 3.5). Unlike ^{15}N-labeling, SILAC can directly isotope-label mammalian cell lines. One of the main advantages of SILAC over many other labeling approaches is that it allows the total lysates from the two SILAC-encoded populations to be combined before any separation or purification steps, which minimizes the often dominant quantitation errors that are otherwise introduced in parallel purification procedures. In SILAC, the relative protein quantity between samples can be directly read-out by measuring the isotopic ratio of peptides by MS after tryptic digestion. Any essential amino acid for the cell culture system under investigation can be used for SILAC encoding, and from the observed mass-difference between the SILAC pair the total number of labeled amino acids in a given peptide sequence can immediately be deduced. Due to the simplicity, ease-of-use and accuracy of SILAC, the approach is becoming very popular and has been validated in many different areas of cell culture-based quantitative proteomics. Our laboratory has recently

established stable isotopic labeling of living model animals like the mouse[98] thereby facilitating proteome comparison in different physiological states. SILAC is particularly well-suited to study post-translational modifications (PTMs) such as phosphorylation changes. We have used SILAC specifically to study diverse cell signaling systems by quantitatively determining the signaling molecules involved and their phosphorylation sites (see refs[82,37] and below).

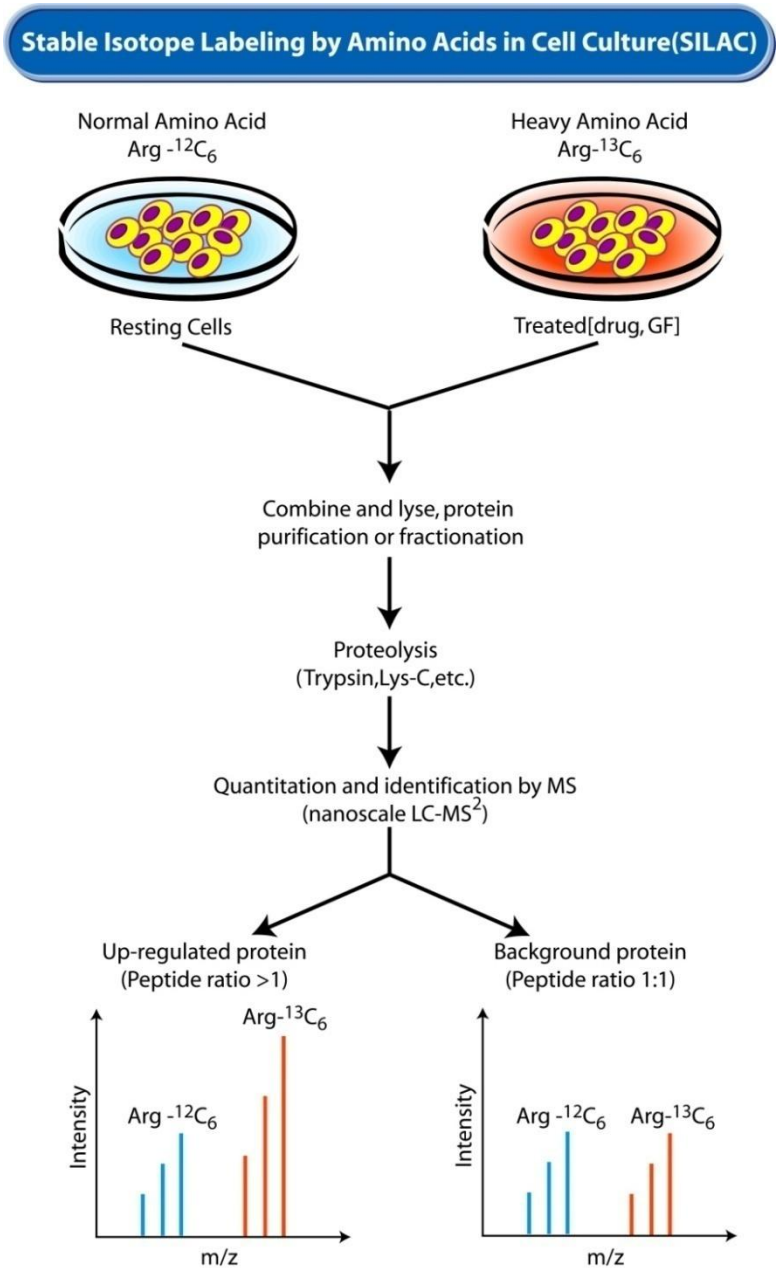


Figure 3.5 Quantitative proteomics by Stable isotope labeling of amino acids in cell culture (SILAC). GF, growth factor.

3.6 Enzymatic isotope labeling (^{18}O)

Peptides can also be isotope labeled via enzymatic reactions. This approach involves specific proteases such as trypsin that incorporate water into the carboxy-termini of peptides during proteolysis[76]. The isotope labeling is accomplished by performing digestion in the presence of stable isotopically enriched water ($H_2{}^{18}O$). Depending on the protease utilized and duration of digestion either one or two ^{18}O atoms can be incorporated into a peptide. Endoproteases such as trypsin, Lys-C and Glu-C are all capable of incorporating two ^{18}O atoms per peptide.

Trypsin digestion in ^{18}O-water therefore results in mass spectra with labeled peptide ion pairs spaced by 2 or 4 Da (^{18}O-labeled vs. the corresponding analogous peptide digested in natural water). Unfortunately, it seems like this methods have some drawbacks that derives from back-exchange of ^{18}O for ^{16}O from solvent during storage and handling of sample[99]. They are therefore not very frequently used.

3.7 Tandem mass tags - iTRAQ

As discussed above, proteins can be quantified relative to each other by comparing the intensities from MS^1 mass spectra of isotope label peptides originating from two different cell states combined in one sample. Alternatively, quantitative information can also be derived from fragment ions in tandem mass spectra. iTRAQ™ is a recently developed protein quantitation method that utilizes isobaric amine specific tags[92] (Figure 3.6).

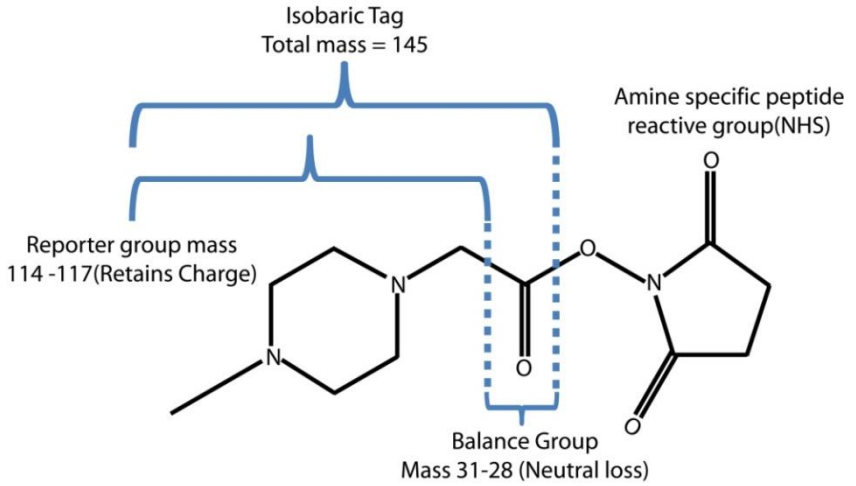


Figure 3.6 iTRAQ reagent

Each tag consists of a reporter and balance group, which is prone to fragmentation, see figure 3.6. The technique is based on chemically tagging the free N-terminus and lysine-ε-amino group on peptides generated from protein digests that have been isolated from cells in different states. The tagged samples are then combined, fractionated by nanoLC and analyzed by tandem MS. In MS^1 spectra the differentially labeled versions of a peptide are indistinguishable. However, in MS/MS spectra each tag generates a unique reporter ion (immonium-like ion). Protein quantitation is then achieved by comparing the intensities of the reporter ions in the MS/MS spectra. There are eight tags available enabling up to eight different conditions to be analyzed in one experiment.

3.8 AQUA and Absolute SILAC for absolute quantitation

In addition to the wealth of isotope labeling strategies that have been developed in proteomics for relative quantitation, Gerber *et al.* have developed a simple approach to measure absolute quantity of peptides and proteins called AQUA[100]. This quantitation technique is based on internal peptide standards, which are synthesized using stable isotope (^{13}C / ^{15}N) labeled amino acids. Known amounts of the peptide or peptides of interest are introduced ("spiked") into the sample and the absolute quantity of the endogenous peptide can readily be derived from the relative ratio observed by MS.

However, due to the costs associated with synthesis of isotopically labeled peptides this strategy has mainly been applied in targeted approaches where a single or few peptides were monitored at the same time. Another drawback of the AQUA strategy is that the peptides are introduced after the digestion step; therefore strictly speaking the absolute amount of peptide at this purification stage rather than in the original sample is determined. Another approach to absolute quantitation is QCAT[101], which is a multiplexed absolute quantitation method using plasmids to synthesize isotope-labeled (^{15}N) proteins. Recently, we developed a new and more robust method for absolute quantitation called "Absolute SILAC"[102]. In this method SILAC-labeled recombinant proteins produced in vivo or in vitro are used as internal standards, which are directly mixed into lysates of cells or tissues - thereby minimizing the systematic errors resulting from differences in sample processing. Successful application of this approach resulted in precise quantitation of

Grb2 copy numbers in HeLa, HepG2 and C2C12 cell lines against the backdrop of whole cell lysate.

3.9 Alternative methods – Quantitation without Stable isotopes

Stable isotope labeling of proteins is not always practical and is relatively expensive. There are several alternative methods for MS quantitation. Although these methods are not as precise and elegant as isotope-labeling approaches they can be used to extract useful information from MS data sets. The most simple and inaccurate approach is called subtractive proteomics, where two samples are compared with MS by subtracting all proteins identified in one sample from all those identified in the other. This is obviously not a very good approach for quantification, because even replicates of the same sample have been shown to have poor overlaps in complex peptide mixture analysis. For example, it has been observed that two replicate MudPIT experiments produced two sets of protein identifications with only 65% overlap[103].

A more elegant way to quantify proteins between a set of samples is to use the extracted ion chromatogram (area under the curve as the peptide elutes from the column), which is in principle available for all peptides if the resolution is sufficiently high. Signal-intensity based quantitation successfully discriminated between true centrosomal proteins and unspecific ones by "protein correlation profiling" of adjacent sucrose fractions[104]. The accuracy of quantitation in this experiment was estimated to be within 30% between peptides derived from the same protein.

Finally, there are of course traditional biochemical methods such as radioactive-labeling, differential 2D-gel dyes and western-blotting approaches. However, these methods are only useful for quantitation of known proteins or require a difficult extra identification step for unknown components.

4. Mass spectrometry data analysis - from ions to protein identification and quantitation

The broad application of proteomics in different biological and medical fields, as well as the diffusion of high-throughput platforms, leads to increasing volumes of available proteomics data requiring efficient algorithms, new data management capabilities and novel analysis, inference and visualization techniques. *Computational proteomics* is an emerging field of biomedical informatics research arising from the demand of high throughput analysis in numerous large-scale experimental proteomics projects[46]. Data analysis in MS based proteomics is a two-tier process where the first few steps are related to the identification and quantitation of proteins (and their post-translationally modified counterparts) followed by the last step of functional proteomics analysis by bioinformatics (Figure 4.1). Computation proteomics mainly deals with the first step of this workflow and has evolved as a separate discipline in itself over the past decade[41].

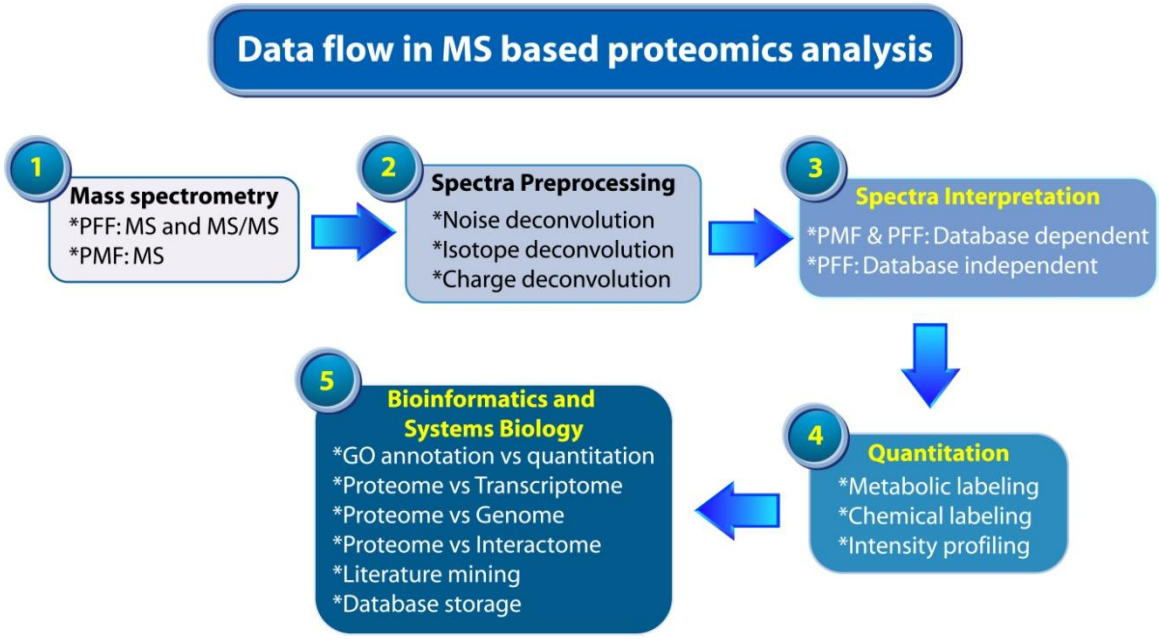


Figure 4.1 Data flow in MS based proteomics(adapted from ref[3]). PFF, peptide fragmentation fingerprinting; PMF, peptide mass fingerprinting

In a proteomics experiment different steps can be identified, each one having one or more parameters of choice: (i) sample preparation, including separation and labeling; (ii) MS experiment, including mass spectrometer choice and configuration; (iii) spectra preprocessing, including spectra signal deconvolution, often performed by the spectrometer software, baseline subtraction, noise removal, dimension reduction, peak extraction, outlier detection; (iv) peptide/protein identification, including database searching, perhaps coupled with *de novo* sequencing or sequence tagging and (v) peptide/protein quantitation, either performed through stable isotope labeling or through intensity profiling. Each of these steps has important implications on the algorithmic and analytical aspects of computational proteomics. The computational facets of all these steps are discussed elaborately in literature and can be found in these reviews[3,105].

4.1 Peptide and Protein identification

Protein identification by peptide sequencing is at the heart of bottom-up or top-down proteomics[42]. The cornerstone in protein identification are peptide database search algorithms such as MASCOT and SEQUEST (based on database and signal processing paradigm)[106,107], which allow automatic matching and scoring of peptide fragment ion spectra against protein sequence databases. This approach for protein identification is also known as **"**searching of uninterpreted mass spectra" or "probability based scoring"**.** When a peptide ion is fragmented in an MS/MS experiment, there are two pieces of information that are known for each peptide. The first is the molecular weight of the peptide and the second is the list of fragment ion masses and their intensities. This information is used by the search algorithm in an attempt to identify the peptide sequence in a given protein database. All entries in the database are digested "*in silico*" using the appropriate enzyme and possible modification masses are added to the mass of the resulting peptides. Every experimentally recorded peptide mass is then compared to all *in-silico* peptides of the same mass, within operator set mass tolerances, and all theoretical peptides with other masses are discarded. Then, the calculated fragment ion masses for all of the peptides in the selected mass range are compared to the experimentally observed fragment ion masses for the peptide - again within operator set mass tolerances - and an peptide score is assigned. The output from the search displays the best matching peptide sequence (if any) in the database for

each MS/MS spectrum and an associated score that indicates how good the match is. In probability based algorithms like MASCOT, the score for an MS/MS match can be given as $-10*\mathrm{Log}_{10}(p)$, where p roughly corresponds to the probability that the observed match between the experimental data and the database sequence is a random event. In the result output the search-software also groups peptides matching the same protein sequence.

Many more bioinformatics approaches to protein identification have been developed in recent years which draw insights from various computational methods, for instance graph theory[108], statistical learning[109] and machine learning approach[110]. Additionally, the identification of post translationally modified peptides and proteins pose formidable computational challenges[111,112], which needs specialized algorithmic treatment [108,113].

4.2 Peptide and Protein Quantitation

Protein quantitation by MS depends on either comparison of labeled (or tagged) peptides or by label-free methods for comparison of identical peptides across different biological states[45,114]. With advances in MS technology and experimental setups for proteomics, a typical proteomic experiment can produce gigabytes of data[115]. Therefore manual analysis of such datasets to extract quantitative information is not possible by proteomics researchers and can only be done in an automated way by using advanced computational algorithms and analytical pipelines. Unlike microarray technology, the lack of standardized and comprehensive quantitation software for MS data has been one of the largest challenges and bottlenecks for proteomics[46,116]. To date various empirical methods like peptide counting[117] and spectral counting has been used[118,119] for protein quantitation. The past few years has witnessed application of machine learning and mathematical and statistical methods for quantitative proteomics[120]. But most of these methods are inherently inaccurate and were developed for low resolution MS data. Therefore they fail to deliver when MS data is highly resolved and fine grained as generated by the latest generation of mass spectrometers. Our laobratory has recently developed a suite of integrated algorithms specifically for high resolution, quantitative MS data based on state-of-the-art data modeling, correlation analysis and graph theory[47]. Though still in its infancy, computational proteomics for protein quantitation is one of the most exciting area for bioinformatics researchers and it is

witnessing rapid developments in computational frameworks, data standardization[121] and software development[3].

5. Bioinformatics for high throughput “omics” sciences

The foundations of modern molecular biology were laid in the early 1950s by pioneering work carried out for elucidating the sequence and biomolecular structures of DNA and proteins[122-124] (Figure 5.1). Since then this field has caused a paradigm shift in biomedical research, and played very instrumental roles in our current understanding of cellular biology.

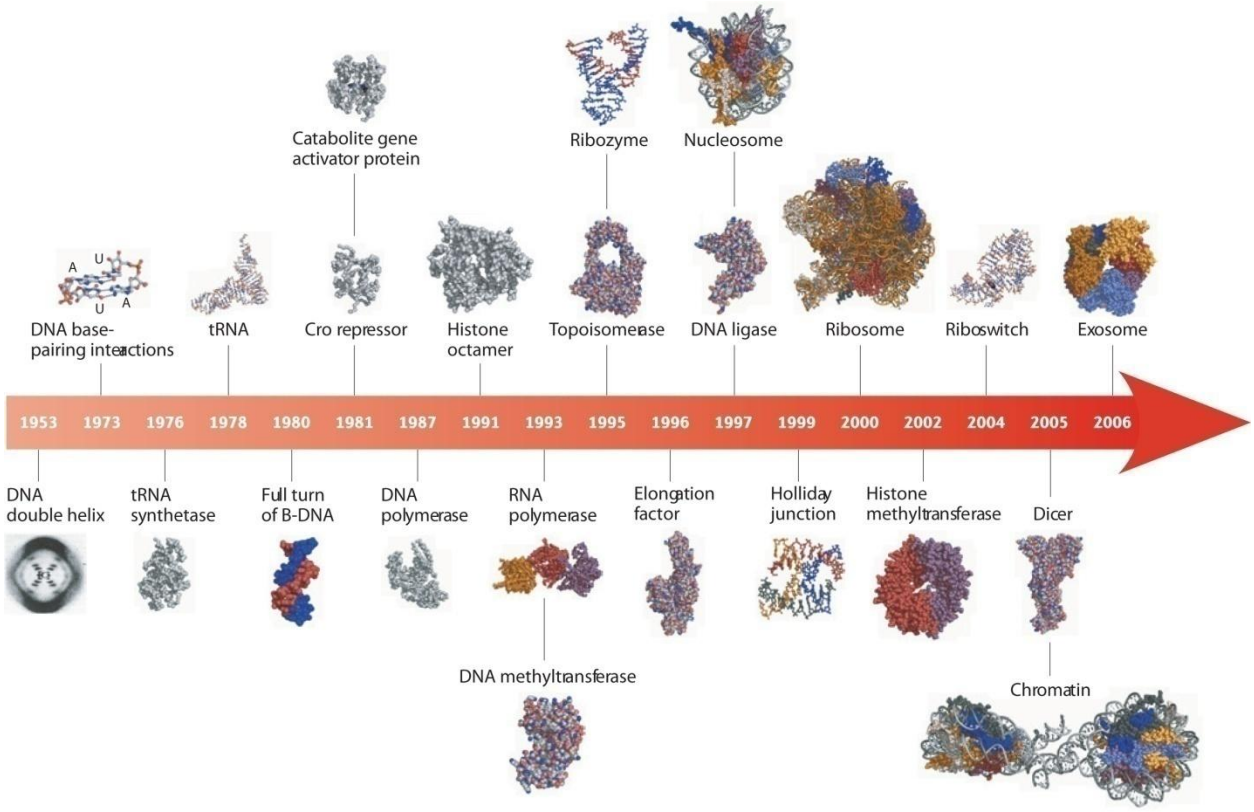


Figure 5.1 Timeline of solved structures of some key biomolecules (adapted from ref[125])**.**

One of the most impressive achievements of molecular biology has been that it bridged the gap between experimental and computational research thereby transforming biology into a quantitative, information driven discipline[126,127]. Some of the most fundamental and scientifically significant questions addressed by molecular biology presented novel, interesting and at times formidable algorithmic problems, and relied heavily on computational resources[4,5]. For instance, the structure of DNA[128], the encoding of genetic information for proteins[129], the factors

governing protein structure[130,131], the structural properties of protein molecules[132-135], the evolution of biochemical pathways[136] and gene regulation[137], and the chemical basis for development[138] all contain seeds of some of the problems that were only possible to address by computation in the ensuing decades. Application of computer technology and computational methods to molecular biology is not a recent trend[5].

Large scale genome sequencing of model organisms and humans at the beginning of the twenty first century[97,139,140] has been a landmark achievement in biomedical sciences that was fueled by extraordinary advances in molecular biology and computer science. This in turn helped in redefining the synergy between biology, mathematics and information sciences, thereby leading to emergence of *bioinformatics* as a scientific discipline. Bioinformatics involves the integration of computers, computational algorithms, software tools, and databases in an effort to address biological questions. Bioinformatics approaches are often used for major initiatives that generate large data sets. In the post genomic era bioinformatics has embedded itself into the very fabric of contemporary biology and become an indispensable part of any ambitious molecular biology enterprise.

5.1 Current state-of-the-art in Bioinformatics

Initial bioinformatics research and applications were primarily focused on analysis of biological sequence data, genome content, and rearrangement, and for prediction of function and structure of macromolecules[141-143](Figure 5.2). With the advent of high throughput "omics" disciplines in the past few decades, it is now feasible to systematically profile a biological system at different levels of molecular and cellular organization, including its epigenome, transcriptome, metabolome, proteome, and interactome[62]. Owing to the largely disparate nature and scale of these data types, newer analysis and research directions have spawned from conventional bioinformatics, for instance - comparative genomics[144], functional genomics[145], network biology[146], and computational proteomics[41]. Moreover, with the further realization that the complete understanding of the physiology of biological systems entails an undertaking of multi-level data integration and analysis, computational systems biology has come into prominence[147].

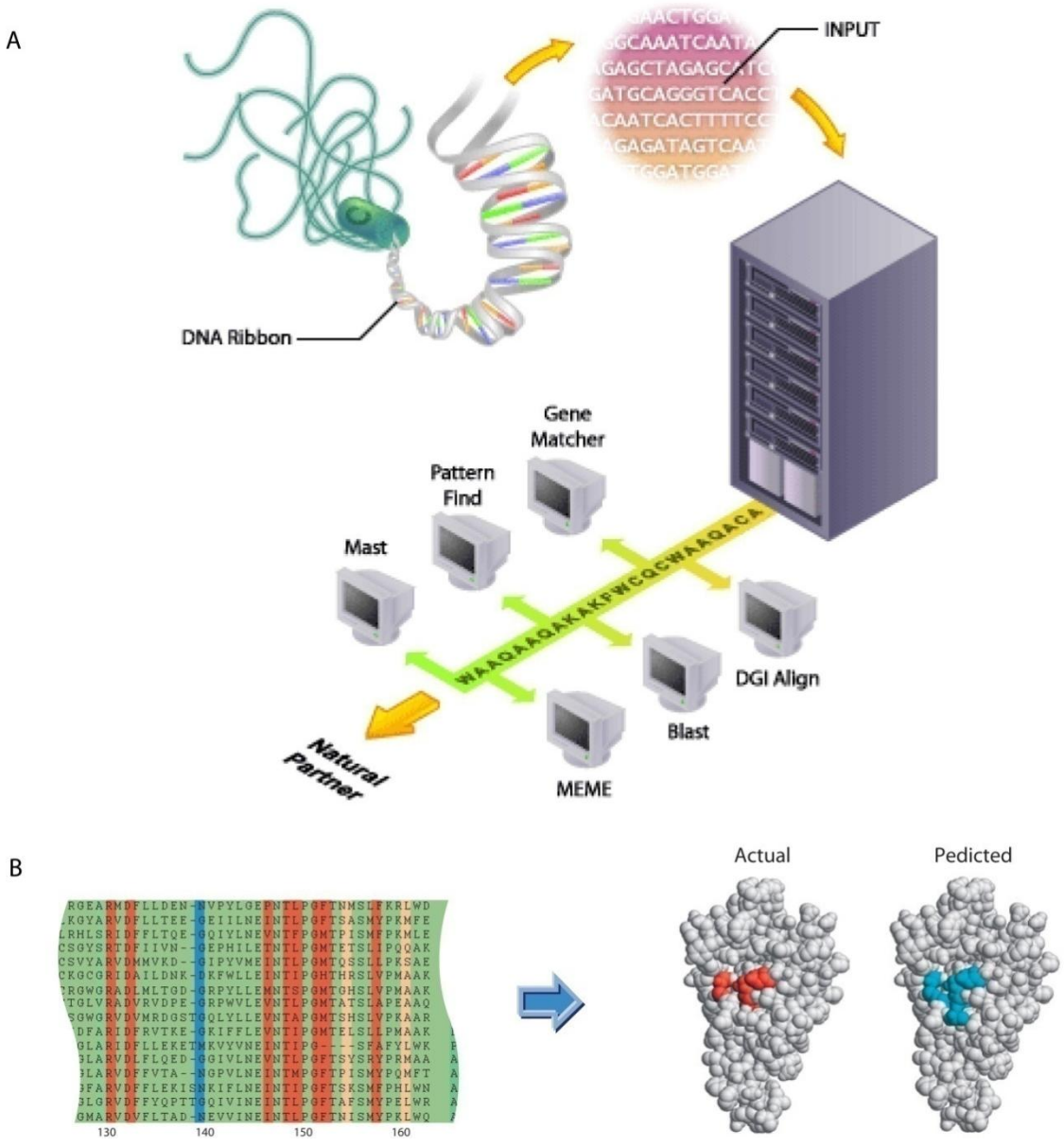


Figure 5.2 The Use of Computers to Process Biological Information.
(A) The wealth of genome sequencing information has required the design of software and the use of computers to process this information. (Image adapted from Joanne Fox article URL: http://bioinformatics.ubc.ca/about/what_is_bioinformatics/). **(B)** The evolutionary trace (ET) method for identifying specificity residues in proteins (adapted from ref[148]). An alignment of related sequences to a protein of interest is created (left panel). The level of conservation of columns in the alignment is calculated (colors represent a sliding scale in which the most conserved residues = red, those with intermediate conservation = green and the least conserved = blue) and conserved positions are mapped onto the structure or sequence of the unknown protein. In the right panel a comparison of the actual substrate-binding residues and predicted residues when the ET method was applied to the ATP-grasp domain of a d‑Ala–d-lactate ligase (Protein Data Bank entry 1EHI). It can be seen that the ET method achieves a close prediction of the substrate-binding site.

The quest for systems level understanding of biological systems has led to evolution of bioinformatics as a multifaceted and complex scientific ecosystem, with active research activities in each of its sub-disciplines. In this section current state-of-the-art in modern bioinformatics research in the context of “omics” disciplines are briefly reviewed, more detailed accounts can be found in the huge body of literature published in last decade[149-152]

5.1.1 Bioinformatics for Gene Expression - Functional Genomics

Microarrays have now permeated literally every field of biology and have found numerous applications in applied and translational research[153-155]. Large-scale microarray studies are also becoming crucial to a new way of conceiving experimental biology[156]. Since the inception of this technology, one of the primary challenges for bioinformatics researchers has been to analyze and mine the high-dimensional dataset generated in a typical microarray experiment[157]. Furthermore, these challenges are accentuated by the diversity of biological questions being investigated by this platform. Microarray data is being used in conjunction with computational algorithms and various machine learning paradigms which finds wider application in drug discovery, basic research and target discovery, biomarker determination, pharmacology, toxicogenomics, development of prognostic tests, population genomics and disease subclass determination. Inferential[158,159] and descriptive statistical[160-162] methods have been successfully applied to microarray data to uncover patterns of gene expression and behavior of genetic markers in diseases, especially cancer[163-166] (Figure 5.3). Reverse engineering of gene networks using gene expression data based on Bayesian statistics (probabilistic models)[167](Figure 5.4) , Boolean networks[168], Relevance Networks[169], and graph theoretic algorithms[170] has gained momentum in recent years[171-173]. The diversified field of analysis of microarray data falls under the umbrella of Microarray Bioinformatics.

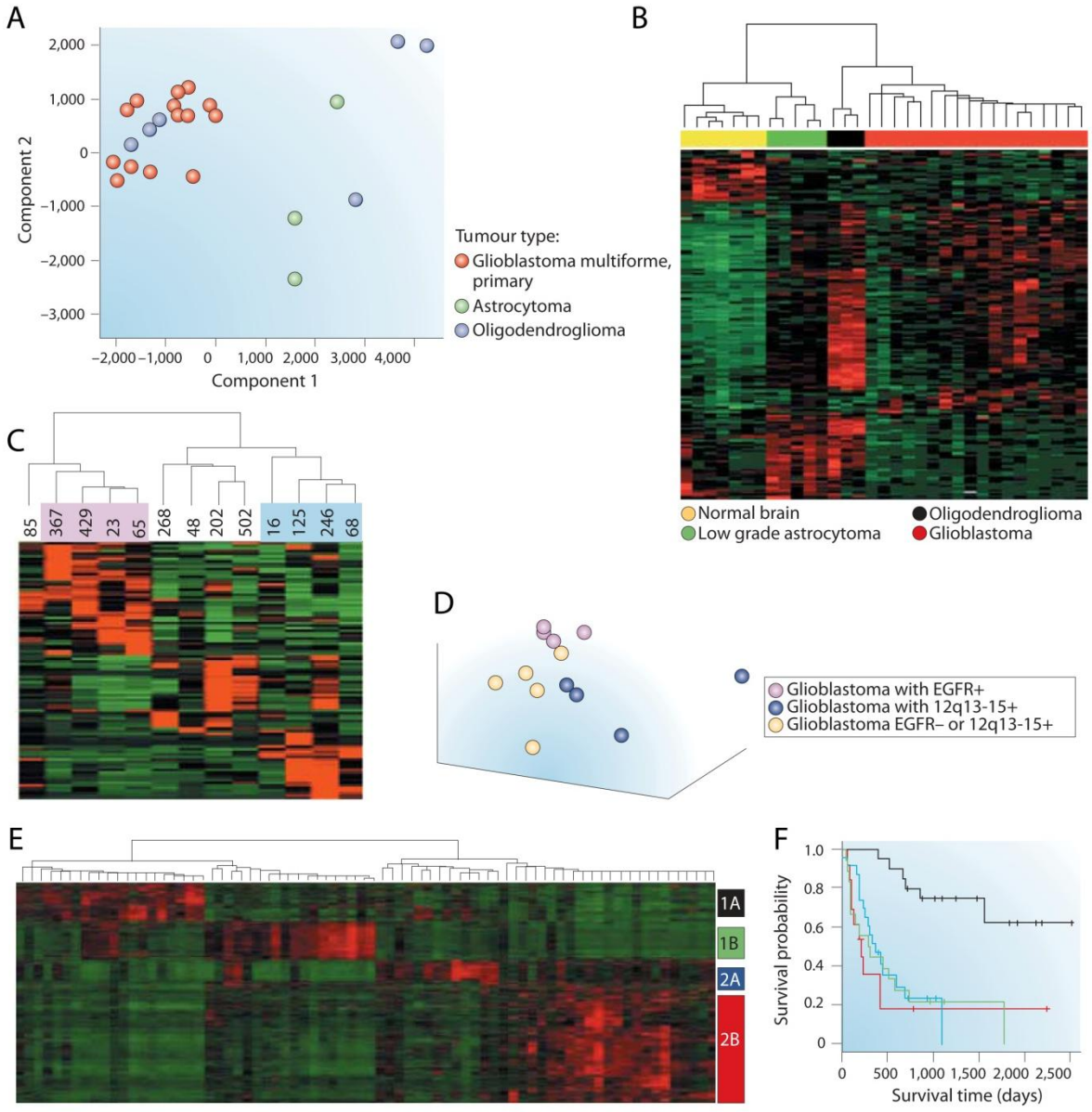


Figure 5.3 DNA-microarray analyses can identify relevant clinical subsets of gliomas (adapted from ref[174]). **(A)** and **(B)** show that different subtypes of gliomas have distinct gene-expression profiles as revealed by multidimensional scaling and hierarchical clustering respectively. **(C)** and **(D)** show identification of molecular subsets of microscopically identical glioblastomas by hierarchical clustering and multidimensional scaling respectively. **(E)** Hierarchical clustering of 85 high-grade glioma samples on the basis of the expression of 595 genes that are highly differentially expressed in patients with relatively good survival times versus those with shorter survival times. Four subsets of patients are detected. **(F)** Kaplan–Meier survival analysis shows that these genes can identify the subset of patients who are most likely to have prolonged survival times (E, cluster 1A, black).

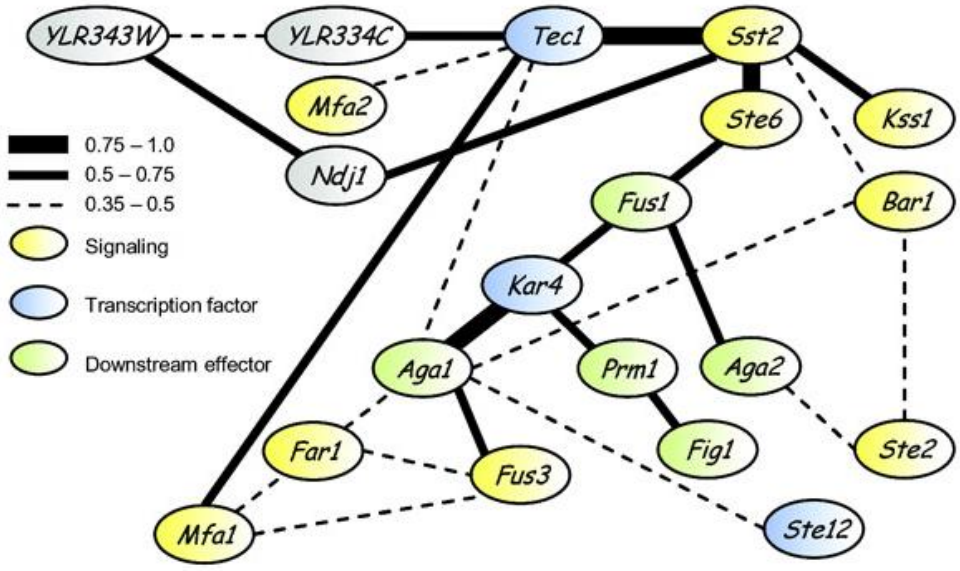


Figure 5.4. Regulatory network architecture learnt from microarray data using reverse engineering algorithms (figure adapted from ref[173]). Shown is an unconstrained acyclic network where each gene can have a different regulator set. This is a fragment of a network learned in the analysis of Pe'er *et al.*[175], performed on Rosetta compendium of expression profiles from budding yeast by Hughes *et al.*[156] . A summary of direct neighbor relations among the genes is shown based on bootstrap estimates. Degrees of confidence are denoted by edge thickness. A sub-network of genes involved in the yeast-mating pathways was automatically identified with high-confidence relations among them. The colors highlight genes with known function in mating, including signal transduction (yellow), transcription factors (blue), and downstream effectors (green).

5.1.2 Bioinformatics of Gene Regulation

The completion of various genome sequencing projects has provided us with the genetic blueprint of genomic organization and structure, thereby leading to many intriguing observations. Foremost among them is the discovery that cells and organisms devote a significant fraction of their DNA to encode *cis*-regulatory programs that control and coordinate gene expression at the transcript level[176]. *Trans*-acting protein and *cis*-regulatory sequences are the principle components of these molecular programs which act in response to a particular cellular context (state) and extra-cellular inputs, and have pivotal role in organismal development and evolution[177]. A complete understanding of this molecular algorithm will have profound impact on biological research that will be essential for gaining insights into development, cellular

responses to environmental and genetic perturbations and the molecular basis of many diseases[178]. While the unraveling of the entire network of gene regulation is a distant goal, promising results in this direction have started to emerge from diverse studies which include bioinformatics approach for identification of *cis*-regulatory sequences (reviewed in ref[179]), and development of *in-silico* mathematical models for gene regulatory networks based on experimental observations[180-182]. With recent interesting discoveries in transcriptional regulation by regulatory RNAs (*mi*RNA,*si*RNA,*pi*RNA) and current appreciation of the role of epigenetics, the bioinformatics research in this domain has gained impetus in computational prediction of the key players of gene regulations including various regulatory RNAs and their targets[183-185] and prediction and modeling of epigenetic information and contexts[186]. The bioinformatics of gene regulation is currently one of the most exciting areas of computational research.

5.1.3 Network Bioinformatics

Cells and organisms have evolved amazingly elaborate and intricate mechanisms to carry out their basic functions. These functions emerge largely as a result of the dynamic interplay between cellular components- *genes, proteins, and metabolites,* which are further interwoven in complex biological networks including webs of protein-protein interactions, regulatory circuits linking transcription factors and *cis*-regulatory targets, signal transduction pathways and metabolic pathways. Recent scientific and technological breakthroughs in high-throughput genomics[187,188] and proteomics[189] are enabling us to discover the molecular programs carried out by these networks. But we are still far from a complete understanding of the design principles, architecture and dynamics of these networks. Promising results in this direction have started to emerge from diverse studies which include mathematical modeling and simulation of pathways, graph-theory analysis of global network structure, application of engineering concepts to network analysis, partitioning of networks into functionally related modules & motifs, and *de novo* design of networks[190-192]. Graph-theory analysis proposed by Barabasi *et al.*[146] provides details about the static topological properties (*degree distribution, clustering coefficient*), internal organization (*hubs, motifs*) and evolution of these networks (Figure 5.5). Application of engineering concepts to biological network analysis has revealed that biological networks share structural principles of modularity, robustness and recurrent circuit elements with their engineering counterparts[193-195].

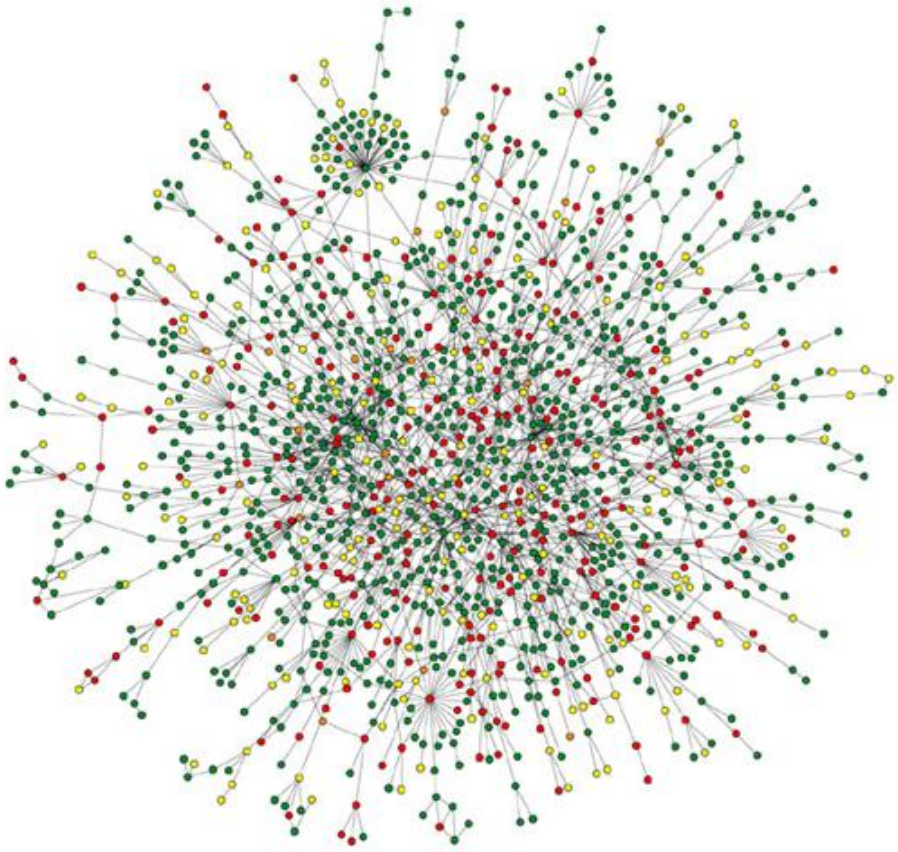
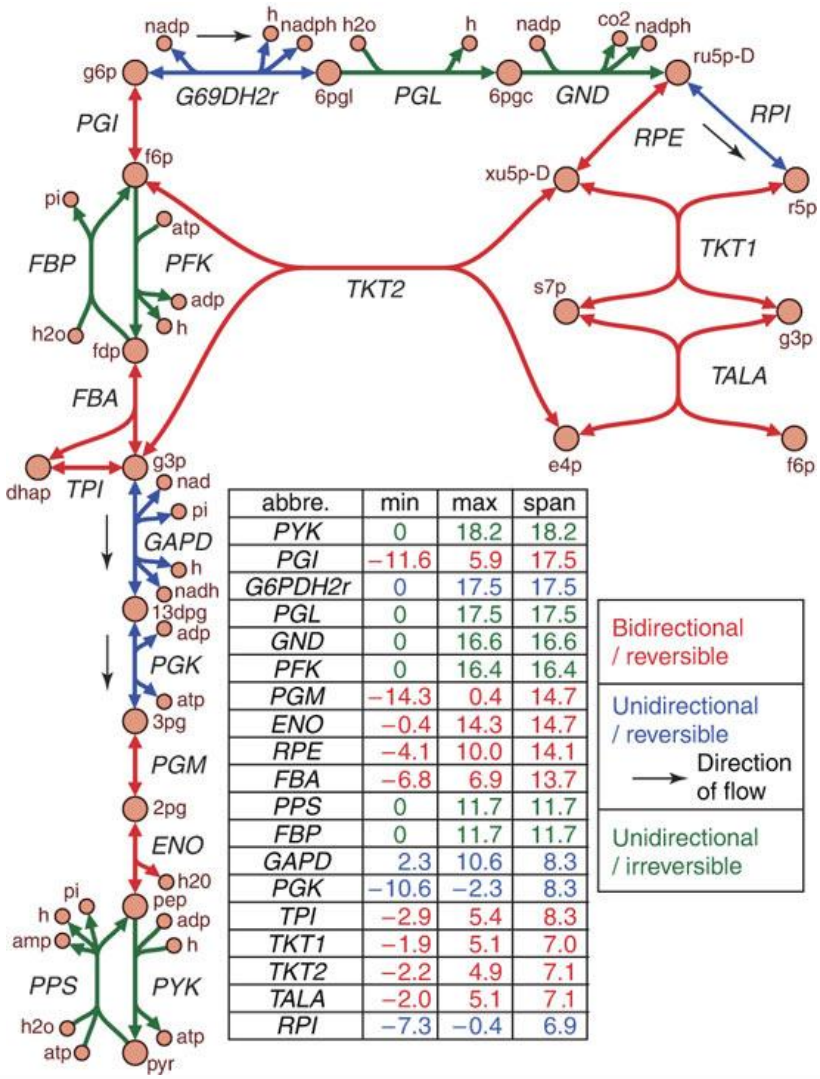


Figure 5.5 A map of protein–protein interactions in *Saccharomyces cerevisiae* (adapted from Jeong *et al.*[196]), which is based on early yeast two-hybrid measurements[197], illustrates that a few highly connected nodes (known as hubs) hold the network together. The largest cluster, which contains ~78% of all proteins, is shown. The color of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown).

Reverse engineering of networks[173,198] based on machine-learning approach has gained momentum in recent years because of the availability of high throughput data, especially from microarrays, yeast two-hybrid screens (Y2H) and chromatin immunoprecipitation (ChIP) experiments. Additionally, approaches dealing with dynamics of gene regulatory and protein-protein interaction networks have been recently reported[199,200]. Integration of phenotypic, drug target and disease annotation information with protein interaction and metabolic networks has led to interesting insights into disease mechanisms, morbidity and drug actions[201-203].

Metabolic network analysis spans another dimension of network bioinformatics whereby specialized analytical methods based on constraint-based approach such as flux variability analysis (FVA) have been popularized by Palsson *et al.*[204-206](Figure 5.6). In recent years network bioinformatics has successfully integrated into the theme of wet biology by providing testable hypothesis to biologists[207, 208].



abbre.	min	max	span
PYK	0	18.2	18.2
PGI	−11.6	5.9	17.5
G6PDH2r	0	17.5	17.5
PGL	0	17.5	17.5
GND	0	16.6	16.6
PFK	0	16.4	16.4
PGM	−14.3	0.4	14.7
ENO	−0.4	14.3	14.7
RPE	−4.1	10.0	14.1
FBA	−6.8	6.9	13.7
PPS	0	11.7	11.7
FBP	0	11.7	11.7
GAPD	2.3	10.6	8.3
PGK	−10.6	−2.3	8.3
TPI	−2.9	5.4	8.3
TKT1	−1.9	5.1	7.0
TKT2	−2.2	4.9	7.1
TALA	−2.0	5.1	7.1
RPI	−7.3	−0.4	6.9

Figure 5.6 A map of some of the reactions in glycolysis and the pentose phosphate pathway of *E. coli* (adapted from *Becker et al.*[209]). Using Flux Variability Analysis (FVA), the minimum (min) and maximum (max) allowable flux values for each reaction were determined using the *E. coli* model *i*JR904. The values shown in the table correspond to the min and max allowable fluxes for each reaction shown in the map when the predicted growth rate is constrained to 90% of the optimal value under glucose-limiting conditions. The results were further characterized by the direction of predicted flux (bidirectional or unidirectional) computed using FVA. The black arrows in the figure show the predicted unidirectional direction of flux for reversible reactions that can potentially operate in either direction.

5.2 Bioinformatics for high-throughput mass-spectrometry proteomics data

Mass spectrometry based proteomics has undergone tremendous advances over the past few years largely due to technological breakthroughs in instrumentation and state-of-the-art innovations in computational proteomics[2]. Proteomics is now a very data intensive discipline that requires extensive analytical and data-mining support, and bioinformatics has thus become a pivotal constituent of this discipline. Proteomics research endeavors can be classified into two broad categories, namely qualitative proteomics and quantitative proteomics. The scope of functional proteomics analysis and the analytical direction that can be taken are largely dependent on the type of dataset generated by these research endeavors, and are discussed next.

5.2.1 Bioinformatics for Qualitative Proteomics

In qualitative proteomics the focus is on enumeration of the proteome constitution of a system of interest - body fluid, cell type, organelle, tissue or an entire organism[50-52,57,210]. Most of the bioinformatics activities therein focus on functional data mining of the dataset to extract the global biological theme underlying the proteome. In recent years genome-wide annotational datasets like gene ontology (GO)[48], protein domain organization (PFAM, InterPro)[211,212], pathways (KEGG)[49], and disease mutations(OMIM) have been successfully used for functional grouping of the proteomes[213]. These annotations can be further used in conjunction with statistical tests to find over/under-represented functional categories[50,52]. Additionally, integration with other high throughput “omics” datasets (microarray, ChIP-chip) and annotations has provided valuable insights into proteome expression and turnover, and their role in disease mechanisms[30,57,214-216]. In the same vein, augmentation of genome annotation[108,217,218], and search for gene models based on high confidence peptide information has shown interesting results in predicting novel genes and splice variants[52,219].

5.2.2 Bioinformatics for Quantitative Proteomics

Quantitative mass-spectrometry (MS) adds another dimension to proteomics studies by providing quantitative data of proteome changes in the cellular states being investigated[45]. This may be either studied on a binary level of protein changes (normal vs. cancer, stimulated with a growth factor vs. non-stimulated) or on multiple levels of protein changes (temporal steps of cell cycle,

maturation of organelle or differentiation of cells). Supervised and non-supervised machine learning approaches have found wider applications in analyzing such datasets[120]. The hierarchical clustering approach has been adopted in various proteomics studies to cluster phospho-proteomic and proteomic datasets across multiple conditions or samples[220]. In the first global analysis of temporal profile of phosphoproteome dynamics, I used fuzzy *k*-means clustering approach to cluster data across four time points[65]. As part of creating an organelle map of mouse liver proteome Foster *et al.* used a supervised approach to assign organelle localization to proteins based on the correlation to organelle specific marker protein dynamics[58]. Dunkley *et al.* used principal component analysis (PCA) methodology to assign organelle localization to protein based on their quantitative data across multiple organelles[221]. Rinner *et al.* employed a combinatorial workflow of label-free mass spectrometry and computational analysis based on an ensemble of *k*-means clustering and expectation maximization (EM), to confidently identify interaction partners in protein complexes[222]. In clinical applications of proteomics, biomarker discovery by comparison of proteome profiles of healthy and disease individuals or cohorts rely heavily on pattern mining approaches – using for instance genetic algorithms[223], and on building classification models for disease predictions[224].

5.3 Prologue to the thesis work study

In chapters 6-9 of the thesis I discuss specific projects illustrating applications of bioinformatics for the functional and systematic analysis of high throughput proteomics data. The projects will be discussed as complete entities with the relevant experimental, computational methods and results. The analysis workflow and computational methods for each of them were developed as part of my PhD study in Department of Proteomics and Signal Transduction at the Max Planck Institute for Biochemistry.

6. In-depth Analysis of the Adipocyte Proteome by Mass Spectrometry and Bioinformatics

This work is included as a manuscript that has been published with the following citation:

Jun Adachi[φ], **Chanchal Kumar**[φ], Yanling Zhang, Matthias Mann

In-depth Analysis of the Adipocyte Proteome by Mass Spectrometry and Bioinformatics
(2007) *Mol Cell Proteomics*; 6(7):1257-1273

[φ] These authors contributed equally to this work

6.1 Introduction

Obesity has become a global health epidemic, which leads to an increased population risk for obesity-related complications such as hypertension, dyslipidemias, type II diabetes mellitus, and cardiovascular diseases associated with the onset of insulin resistance[225,226]. Studies in the last few years have transformed our thinking about the function of adipocytes (fat cells) in physiology in general and obesity in particular[227]. They are no longer regarded just as a passive depot for storing excess energy in the form of triglyceride but as endocrine cells that actively regulate the pathways responsible for energy balance by the secretion of various bioactive substances termed adipocytokines. Furthermore recent research has highlighted the lipid droplet as a dynamic and actively regulated organelle[228,229]. To elucidate the pleiotropic functions of the adipocyte, several proteomics studies had been conducted. However, a large scale proteomics study of adipocytes has not been reported before and could serve as a useful resource for fundamental biomedical research. These considerations as well as our interest in insulin signaling and the metabolic syndrome prompted us to perform an in-depth proteomics analysis of the cellular and organellar proteome of adipocytes. We used a combination of one-dimensional gel electrophoresis and on-line electrospray tandem mass spectrometry with biochemical procedures for sub-fractionation of the cellular proteome (Figure 6.1). State of the art protein identification technology recently developed in our laboratory, involving a linear ion trap (LTQ) -FTICR mass spectrometer with very high mass accuracy[84], allowed us to identify more than 3,200 proteins with extremely stringent identification criteria.

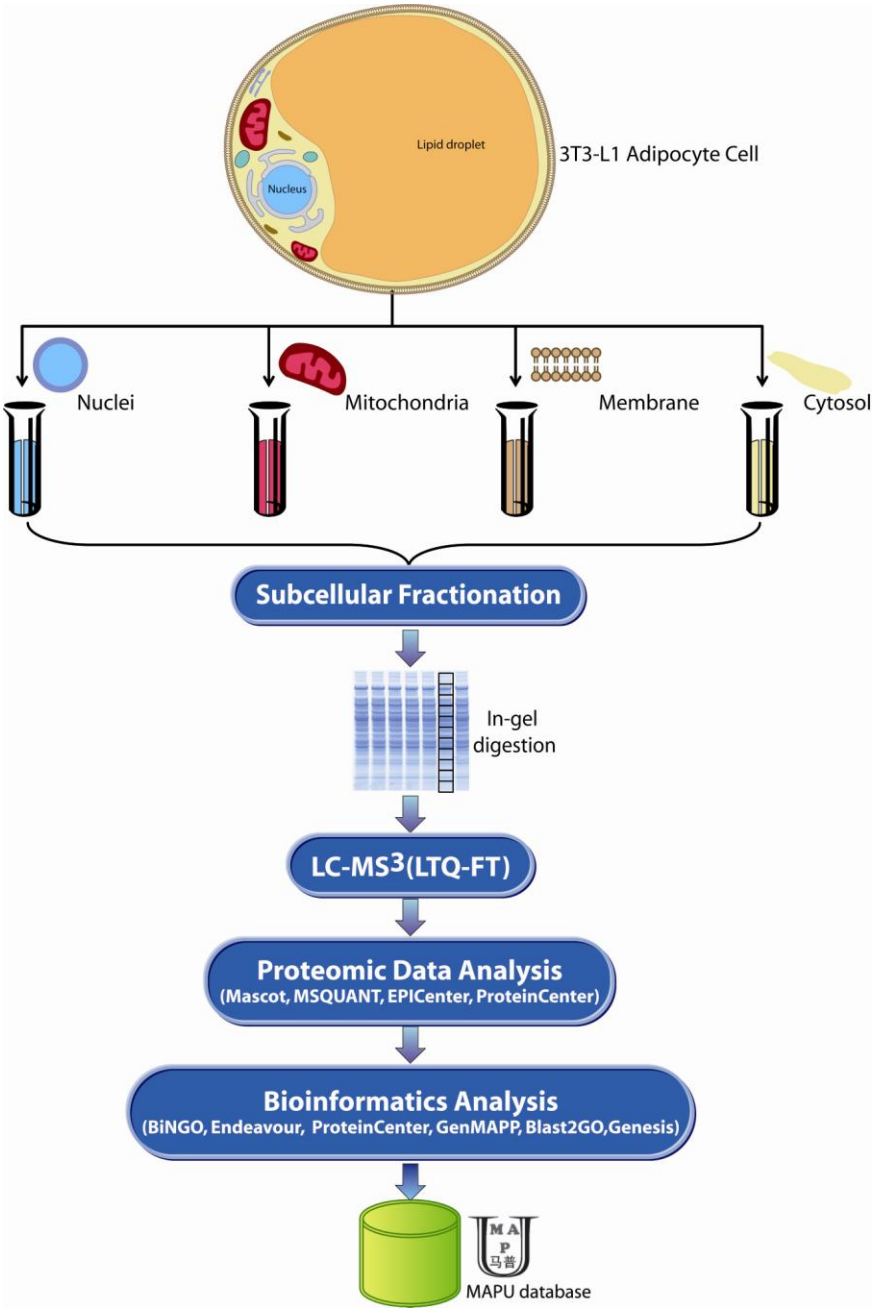


Figure 6.1 An overview of the experimental and bioinformatics procedures used for analysis of the adipocyte proteome. MAPU database: Max-Planck Unified Proteome database.

Extensive bioinformatics analysis and comparison with transcriptome data revealed several layers of information related to the adipocyte proteome that were in turn mapped to an ensemble of biological processes, functions, and pathways. Additionally by using a systemic protein

prioritization methodology described recently[8] we predicted candidate proteins hitherto not known to be involved in insulin-dependent vesicular trafficking.

6.2 Materials and Methods

6.2.1 Cell culture

The maintenance and differentiation of mouse 3T3-L1 preadipocytes were essentially as described[230,231]. Briefly, two days after the cells reached confluence, a cocktail of 0.5 mM 3-isobutyl-1-methyxanthine (mix; Sigma), 1 µM dexamethasone (Sigma), and 167 nM insulin (Sigma) was supplied during culture (day 0). After 48 h (day 2), the cocktail was replaced with only 167 nM insulin. An additional 48 h (day 4) later, insulin was withdrawn, and the medium was changed every second day. Fat accumulation was measured by Oil Red O staining and differentiated cells at day 9 were used for further experiments.

6.2.2 Subcellular fractionation and western blotting

Differential centrifugation was used to fractionate adipocytes as described previously [232,233]. Briefly, cells were sheared by 10 passages through a 25-gauge needle and centrifuged at 1,000 *g* for 10 min at 4 °C. The supernatant was called crude cytoplasm and served as the source of cytosol, mitochondria and membranes. The pellet contained nuclei. (1) To obtain pure nuclei, the pellet was resuspended in sucrose-TKM buffer and overlaid in the 1.6 - 2.3M sucrose gradient. After centrifugation at 160,000 *g* for 1 h (Sorvall surespin 630 rotor), the nuclei settled at the 2.1-2.3 M interface. This nuclear layer was isolated, diluted and centrifuged at 2,700 *g* for 10 min. The final nuclear pellet was resuspended in 200 µl of HES buffer. (2) To obtain pure mitochondria, the crude cytoplasm was centrifuged at 16,000 *g* for 1 h. The resulting pellet was suspended in TES buffer and centrifuged at 16,000 *g* for 20 min. The pellet was suspended, while the supernatant served as the microsomal and cytosolic fraction. This supernatant was overlaid with 0.58 - 1.55 M sucrose step gradient and centrifuged at 160,000 *g* for 1 h. The mitochondrial layer at the 1.29-1.55 M interface were isolated, diluted in 15 ml of 0.25 M HES buffer and centrifuged at 16,000 *g* for 20 min. The purified mitochondrial pellet was resuspended in 300 µl of HES buffer. (3) To obtain pure membrane fraction, the post-mitochondrial

cytoplasmic fraction was centrifuged at 160,000 *g* for 1 h. The supernatant served as the cytosolic fraction. The pellet was resuspended in HES buffer and centrifuged again at 160,000 *g* for 1 h. The purified membrane pellet was resuspended in 200 µl of HES buffer. (4) To enrich proteins in the cytosolic fraction, Centriprep YM-3 membrane concentrators (Millipore, Billerica, MA) were used. Total protein in all fractions was then quantified using the Coomassie Protein Assay Kit (Pierce, Rockford, IL) and stored at -80 ºC. Equal protein amounts (7 µg) from each of the subcellular fractions were loaded onto 10 and 15% SDS-polyacrylamide gels, electrophoresed, transferred to nitrocellulose membranes, and immunoblotted with antibodies against histone H3 (Cell signaling technology, Beverly, MA), cytochrome C (BD Pharmingen, San Diego, CA), insulin receptor β chain (Santacruz Biotechnology, Santacruz, CA) and MEK1 (upstate, Lake Placid, NY), followed by HRP-conjugated secondary antibodies. The membranes were subjected to chemiluminescent detection according to manufacturer's instructions (ECL, GE Healthcare, Piscataway, NJ).

6.2.3 1D-SDS-PAGE and in-gel digest

Proteins (100, 150, 150 and 150 µg) from each of the subcellular fractions (nuclear, mitochondrial, membrane and cytosol fractions, respectively) were separated by one dimensional SDS-PAGE, using NuPage® Novex Bis-Tris gels and NuPage® MES SDS running buffer (Invitrogen, Carlsbad, CA) according to instructions of the manufacturer. The gel was stained with Coomassie using Colloidal Blue Staining Kit (Invitrogen). Protein bands were excised and subjected to in-gel tryptic digestion essentially as described[234]. Briefly, the gel pieces were destained and washed, and, after dithiothreitol reduction and iodoacetamide alkylation, the proteins were digested with porcine trypsin (modified sequencing grade; Promega, Madison, WI) overnight at 37 ºC. The resulting tryptic peptides were extracted from the gel pieces with 30% acetonitrile, 0.3% trifluoroacetic acid and 100% acetonitrile. The extracts were evaporated in a vacuum centrifuge to remove organic solvent, then desalted and concentrated on reversed-phase C18 StageTips as previously described[235].

6.2.4 Nanoflow LC- MS^2 or MS^3

All nanoflow LC-MS/MS and MS/MS/MS experiments were performed basically as described previously[84,236]. All digested peptide mixtures were separated by online reversed-phase (RP)

nanoscale capillary liquid chromatography (nanoLC) and analyzed by electrospray mass spectrometry (ES MS/MS and MS/MS/MS). The experiments were performed on an Agilent 1100 nanoflow system connected to an LTQ-FTICR mass spectrometer (Thermo Electron, Bremen, Germany) equipped with a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark). Binding and chromatographic separation of the peptides took place in a 15 cm fused silica emitter (75 µm inner diameter) in-house packed with reversed-phase ReproSil-Pur C_{18}-AQ 3 µm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). Peptide mixtures were injected onto the column with a flow of 500 nl/min and subsequently eluted with a flow of 250 nl/min from 10% to 64% acetonitrile in 0.5% acetic acid, in a 105 min gradient. Data were acquired in data-dependent mode using Xcalibur software. The precursor ion scan MS spectra (m/z 300–1575) were acquired in the FT ICR with resolution R = 25000 at m/z 400 (the number of accumulated ions 5×10^6). The three most intensive ions were isolated and fragmented in linear ion trap by collisionally induced dissociation using 3×10^4 accumulated ions. They were simultaneously scanned by FT ICR-selected ion monitoring with 10-Da mass range, R = 50000 and 5×10^4 accumulated ions for even more accurate molecular mass measurements. For MS^3, most intense ion with $m/z > 300$ in each MS^2 spectra were further isolated and fragmented. In data-dependent LC/MS^2 experiments dynamic exclusion was used with 30-s exclusion duration.

6.2.5 Proteomic data analysis

Proteins were identified via automated database search (Mascot; Matrix Science, London, United Kingdom) of all tandem mass spectra against an in-house curated version of the Mouse International Protein Index protein sequence database (IPI, versions 3.07) containing all mouse protein entries from Swiss-Prot, TrEMBL, RefSeq and Ensembl as well as frequently observed contaminants (porcine trypsin, achromobacter lyticus lysyl endopeptidase and human keratins). A ‘decoy database’ was prepared by reversing the sequence of each entry and appending this database to the forward database. Carbamidomethyl cysteine was set as fixed modification, and oxidized methionine, protein N-acetylation, N-pyroglutamate and deamidation of asparagine and glutamine were searched as variable modifications. Initial mass tolerances for protein identification on MS peaks were 5 ppm and on MS/MS peaks were 0.5 Da. Two “missed cleavages” was allowed. The instrument setting for the Mascot search was specified as “ESI-Trap”. Peptide identification information was extracted from the Mascot result file into

EPICenter (Proxeon Biosystems)[237]. Besides the standard search engine results used for peptide assignment (score, expected versus calculated fragment ions, delta mass), additional empirical information was computed by the EPICenter peptide validation module to assist in the assignment[237]. Peptides satisfying the following four criteria were accepted for identification. 1) Peptides for which MS^2 score were above the 99^{th} percentile of significance (Mascot score > 32); 2) Fully tryptic peptides with sequence length 7 or longer; 3) Peptides for which delta scores (the difference in score between 1st and 2nd scoring peptide) were at least 5.0; and 4) Peptides for which y-ion or b-ion score was at least 50.0. The Mascot result file was also imported into MSQuant, open source software available at http://msquant.sourceforge.net, and the MS^3 score was calculated automatically. Finally the protein identification list was created by accepting the peptides (which passed our criterion mentioned before) and consolidating them as per the following method. Proteins with at least two peptides and a MS^2 score of at least 64 were counted as identified proteins. This protein identification criterion corresponds to the confidence of $p = 0.0001$ if both peptide identifications are considered independent. For proteins identified by a single peptide, we required the presence of an MS^3 spectrum and a combined score for MS^2 and MS^3 of above 52 which corresponding to a level of false positives of $p = 0.0001$. EPICenter automatically assigns identified peptides to proteins and organizes all proteins with shared peptides into a single group (protein group). EPICenter selects the protein with most of the peptides an anchor protein and marks proteins that are identified by at least one distinct and separate peptide as conclusively identified proteins. We counted proteins as *identified* only when a protein conclusively identified as described above or a protein group consisted of only isoforms or overlapping database entries.

6.2.6 Enrichment analysis of Gene Ontology (GO) categories

BinGO[238] - the Cytoscape[239] plugin for finding statistically over or under represented Gene Ontology (GO) categories, was used for the enrichment analysis of our liver proteome dataset. The 3T3-L1 proteome dataset was compared against a reference set of complete mouse proteome (IPI mouse v 3.07) GO annotations. A custom GO ontology file for the reference set of the whole IPI version 3.13 mouse dataset was created by extracting the GO annotations available for mouse IPI IDs from EBI GOA Mouse 22.0 (containing 32,776 protein annotations). The analysis was done using the "HyperGeometric test" and we selected all GO terms which were significant with

$p < 0.001$, after correcting for multiple term testing by "Benjamini & Hochberg False Discovery Rate". The analysis was done separately for GO biological process and molecular function categories, and fold enrichment for every over-represented term in the two GO categories were calculated. In the following we discuss the fold enrichment calculation for GO biological process and the same procedure applies for the molecular function category. Suppose the set of over-represented biological process GO terms is called B_{GO}. For each term $B_{GO}[\ ]$ in set B_{GO} the fold enrichment measure was calculated by following formula: $fold(B_{GO}[\ ]) = \frac{\%Adipocyte(B_{GO}[\ ])}{\%IPIMouse(B_{GO}[\ ])}$.

Where $\%Adipocyte(B_{GO}[\ ]) = \frac{Count(Adipocyte\_annotated\_with\_B_{GO}[\ ])}{Count(Adipocyte\_annotated\_in\_GO\_Bio\_process)}$ and,

$\%IPIMouse(B_{GO}[\ ]) = \frac{Count(IPIMouse\_annotated\_with\_B_{GO}[\ ])}{Count(IPIMouse\_annotated\_in\_GO\_Bio\_process)}$.

6.2.7 InterPro domain enrichment for insights into protein function

InterPro (release 13.0) annotations were used for finding statistically enriched protein domains in our dataset. We used the Cytoscape[239] plugin BiNGO[238] for domain enrichment analysis. For domain enrichment we needed three components: 1) the test dataset of identified 3T3-L1 proteome; 2)The InterPro ontology which was built by parsing the "interpro.xml" file (for release 13.0) available at ftp://ftp.ebi.ac.uk/pub/databases/interpro/ using in-house scripts; and 3) the reference set of InterPro annotation for the complete mouse proteome, which was created by parsing the "ipi.MOUSE.IPC" file that contains all the InterPro matches for IPI mouse 3.19 database, available at ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ as part of IPI 3.19 release.

The test set of 3T3-L1 proteome was compared against the InterPro annotations of the IPI mouse reference set using the custom InterPro ontology as the reference. The enrichment analysis was done using the "HyperGeometric test" and we selected all InterPro domains which were significant with $p < 0.001$, after correcting for multiple term testing by "Benjamini & Hochberg False Discovery Rate". The set of over-represented InterPro terms is called I_{enrich}.

For each term $I_{enrich}[\ ]$ in set I_{enrich} the fold enrichment measure was calculated by the following formula: $fold(I_{enrich}[\ ]) = \frac{\%Adipocyte(I_{enrich}[\ ])}{\%IPIMouse(I_{enrich}[\ ])}$.

Where $\%Adipocyte(I_{enrich}) = \frac{Count(Adipocyte\_annotated\_with\_I_{enrich})}{Count(Adipocyte\_annotated\_in\_Interpro)}$ and, $\%IPIMouse(I_{enrich}) = \frac{Count(IPIMouse\_annotated\_with\_I_{enrich})}{Count(IPIMouse\_annotated\_in\_Interpro)}$. Subsequently these enriched InterPro domains were grouped in functional categories as per their representative biological functions.

6.2.8 Proteome mRNA concordance analysis for 3T3-L1 adipocytes

To estimate the depth of the proteome we covered in our survey, we compared our identified proteome list with the microarray dataset for normal 3T3-L1 adipocyte available at the DGAP site (http://www.diabetesgenome.org/chipperdb/expt.cgi?id=60). The available dataset is in triplicates for Affymetrix MG_U74A, B and C array types. We used only MG_U74A (containing 12,654 probe sets) data for analysis because of practical limitations in data analysis. This was primarily because of the difference among the probe-sets for these 3 platforms, which makes comparison difficult. We used "DCHIP" software for the analysis of the microarray data (http://biosun1.harvard.edu/complab/dchip/). The analysis was carried out in two steps. In the first step we estimated the basal expression of the 3T3-L1 adipocyte transcriptome and in second we mapped our 3T3-L1 adipocyte proteome dataset on the trascriptome data. The expression of probe-sets on the triplicates was calculated using "PM-only model", and was further normalized using "invariant set normalization" method[165]. The expression values were then converted to log2 scale. The data was further filtered based on the Present (P) verses Absent (A) call percentage which are widely accepted measure of micro array data quality. We used a criterion of 66% P call for accepting a probe set as expressed i.e. a probe-set was accepted if it had a P call in two out of three samples. Only 5,148 probe sets out of 12,654 met this criterion and they were taken as surrogate for basal 3T3-L1 mRNA expression. Subsequently we mapped our proteome list on the estimated basal expression set. We used Ensembl MartView (http://www.ensembl.org/Multi/martview) release 39 and the "Mus musculus genes NCBIM36" dataset, to map adipocyte IPI IDs to their Ensembl counterparts. The Ensembl IDs were then used to retrieve the MG_U74A probe-sets ids using Affymetrix's NetAffx Analysis Center (http://www.affymetrix.com/analysis/index.affx). Thus we could map 3,287 IPI identifiers to 2,113 MG_U74A probe-sets. Finally the overlap of the adipocyte basal (5,148) probe-sets and

our survey (2,113) probe-sets was calculated. This gave us a final number of 1,755 probe-sets which were found in both datasets. Hence these 1,755 probe-set data were regarded as the genes which we could identify and remaining 3,393 probe sets were used as the not-identified set. We use this consolidated information to calculate the average mRNA expression for the identified verses non-identified proteome using the expression levels of 5,148 probe-sets.

6.2.9 Protein prioritization analysis

The recently reported software application for the computational prioritization of genes - Endeavour[8] was used for protein prioritization. 3T3-L1 proteins (IPI IDs) were mapped to human orthologs (Ensembl Gene ids) using Ensembl MartView release 39 and the "Mus musculus genes NCBIM36" dataset. In total 2,990 IPI protein ids were successfully mapped to human Ensemble gene ids. The training set ($S_{Training}$) was created by choosing 29 genes involved in vesicular trafficking in insulin signaling pathway as shown in Figure 6C. The mapped 3T3-L1 proteome Ensembl list was taken as candidate test set (S_{Test}). The following data sources were used for ranking: 1) literature (abstracts in EntrezGene); 2) functional annotation (Gene Ontology); 3) microarray expression (Atlas gene expression); 4) EST expression (EST data from Ensembl); 5) protein domains (InterPro); 6) pathway membership from KEGG(Kyoto Encyclopedia of Genes and Genomes); 7) cis-regulatory modules (TOUCAN); and 8) sequence similarity (BLAST) data. The model for "vesicular trafficking" genes was created in Endeavour using the above mentioned data sources. Finally the candidate test set (S_{Test}) genes were ranked for their putative role in vesicular trafficking in insulin pathway by measuring their similarity with genes in training set ($S_{Training}$).

6.2.10 Annotating hypothetical proteins using orthology based annotation transfer

To assign putative functions to 335 "hypothetical protein" IDs we used Blast2GO[240] tool (http://www.blast2go.de/) which assigns GO annotation to an unknown protein based on its sequence similarity (orthology) to other protein sequences in a pre-selected database. The GO annotations are assigned based on a four-tier annotation mapping procedure as described in the original paper (ref[240]). We used Swiss-Prot database for this analysis as it serves as the most comprehensive experimentally validated protein database.

6.2.11 Hierarchical clustering of cellular compartment profiles of the adipocyte proteome

A cellular compartment distribution matrix was created for the 3,287 IPI ids that we identify in our analysis. Briefly suppose ***C*** is the 3,287 by 4 cellular compartment matrix corresponding to 3,287 proteins and 4 compartments (nuclei, mitochondria, membrane, cytosol). For a particular IPI say *i*, $i \in [1, 3287]$ and a particular compartment column *j*, $j \in [1, 4]$ if the IPI was observed with *k* unique peptides in the compartment, we place ***C***[*i, j*] = *k*. Else if the IPI *i* was not observed in compartment column *j* then we place ***C***[*i, j*] = 0. This 3287 by 4 matrix was called cellular compartment distribution matrix. The matrix was further converted to a probability distribution matrix C_{prob} of same dimensions (3287 by 4) with each element calculated by the following formula: $C_{prob}[i, j] = \frac{C[i, j]}{\sum_{j=1}^{4} C[i, j]}$. The matrix C_{prob} was then used for one dimensional hierarchical clustering using Genesis[241] software. The distance metric used was "Euclidean" and the clustering was done using "Average Linkage clustering" technique. Subsequently the data from earlier large scale studies[242,243] was overlaid on the clustered dendrogram to ascertain the proteome concordance and depth of our fractionation and sub-cellular identification. Also, the available GO cellular component terms corresponding to the 4 compartments were extracted for the 3,287 IPIs and overlaid on the clustered dendrogram.

6.2.12 Pathway mapping of identified proteins in subcellular compartments

We used the recently developed functional mapping tool GenMapp version 2.1 (http://www.genmapp.org)[244] to map our 3T3-L1 adipocyte dataset on publicly available mouse MAPPs. IPI IDs were mapped to their MGI counterparts using IPI Protein cross reference information as available for IPI mouse version 3.13. Overall 3,124 IPI IDs (95.0% of the total) could be mapped to their respective MGI ids. Subsequently we created a compartment wide list for the mapped MGI ids based on the presence/absence of a particular protein in either of four compartments and the data was mapped to latest available mouse MAPPs.

6.3 Results

6.3.1 High confidence protein identification of mouse adipocyte organelles

In order to reduce the complexity of the proteome and obtain compartment specific information in adipocytes, we performed differential ultracentrifugation from differentiated 3T3-L1 adipocyte cells. The proteome of four compartments (nuclei, mitochondria, membrane and cytosol) were examined using the workflow depicted in Figure 6.1. As shown in Figure 6.2, purity of subcellular compartments was excellent as visualized by Western blots of organellar markers across the four fractions.

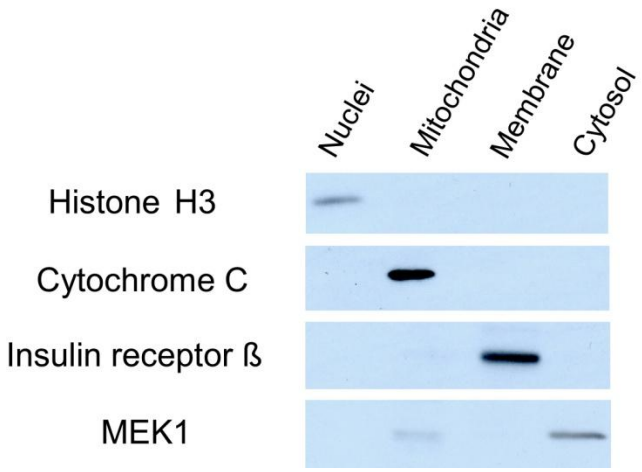


Figure 6.2 Distribution of known organelle markers in subcellular fractions isolated from 3T3-L1 adipocytes

To further reduce sample complexity and dynamic range in protein abundance levels, proteins in each compartment were separated on 1D-SDS-PAGE, and 11 or 12 bands were excised and subjected to in-gel tryptic digestion. In total, 45 fractions were analyzed by liquid chromatography (LC) on-line coupled to electrospray mass spectrometry. We employed a hybrid mass spectrometer consisting of a linear ion trap coupled to a high resolution Ion Cyclotron Resonance (LTQ-FTICR) instrument. The mass spectrometer was programmed to perform survey scans of the whole peptide mass range, select the three most abundant peptide signals and perform narrow range, selected ion monitoring (SIM) scans for high mass accuracy measurements. Simultaneously with the SIM scans, the linear ion trap fragmented the peptide, obtained an MS/MS spectrum and further isolated and fragmented the most abundant peak in the

MS/MS mass spectrum to yield the MS^3 spectrum[236]. Figure 6.3A shows a mass spectrum (MS) of eluting peptides (see ref[42] for an introduction to peptide sequencing). A selected peptide was measured in SIM mode and fragmented (MS^2) (Figure 6.3B). The most intense fragment in the MS^2 spectrum was selected for the second round of fragmentation (Figure 6.3C). As can be seen in the figure, high mass accuracy, low background level and additional peptide sequence information obtained from MS^3 spectra yield high confidence peptide identification. Total cycle time for the analysis described above was approximately five seconds. To obtain the protein 'parts list' of adipocytes, high confidence protein identification and reporting were essential. We applied a stringent multistep filter to minimize false-positive identifications while maintaining favorable detection of lower-abundance and lower-molecular weight proteins (See Section 6.2.5). In addition to standard search engine results i.e. the Mascot probability score[106], MS^3 score and additional empirical information (y-ion and b-ion score, number of sibling peptides and proline score)[237] were employed for peptide and protein identification. Proteins were identified with criteria corresponding to an estimated probability of false positives of $p = 0.0001$. We also performed a decoy database search[245] to test the experimental level of false positive rate in our data set. After applying the stringent criteria mentioned above, we found no false positives for protein identified with two or more peptides and only two false positive protein hits with one peptide. These results indicate a false-positive identification rate of 2/3287 or 0.06%, at the protein level. Thus we conclude that our data set contains no or very few false positive identifications. Determining the identities of proteins from sequenced peptides is complicated because the same peptide sequence can be present in multiple different proteins or protein isoforms[246]. Standard search engines such as Mascot report proteins even when they do not have distinct peptides with their sequence specific to these proteins. Thus, sharing information on identified peptides was

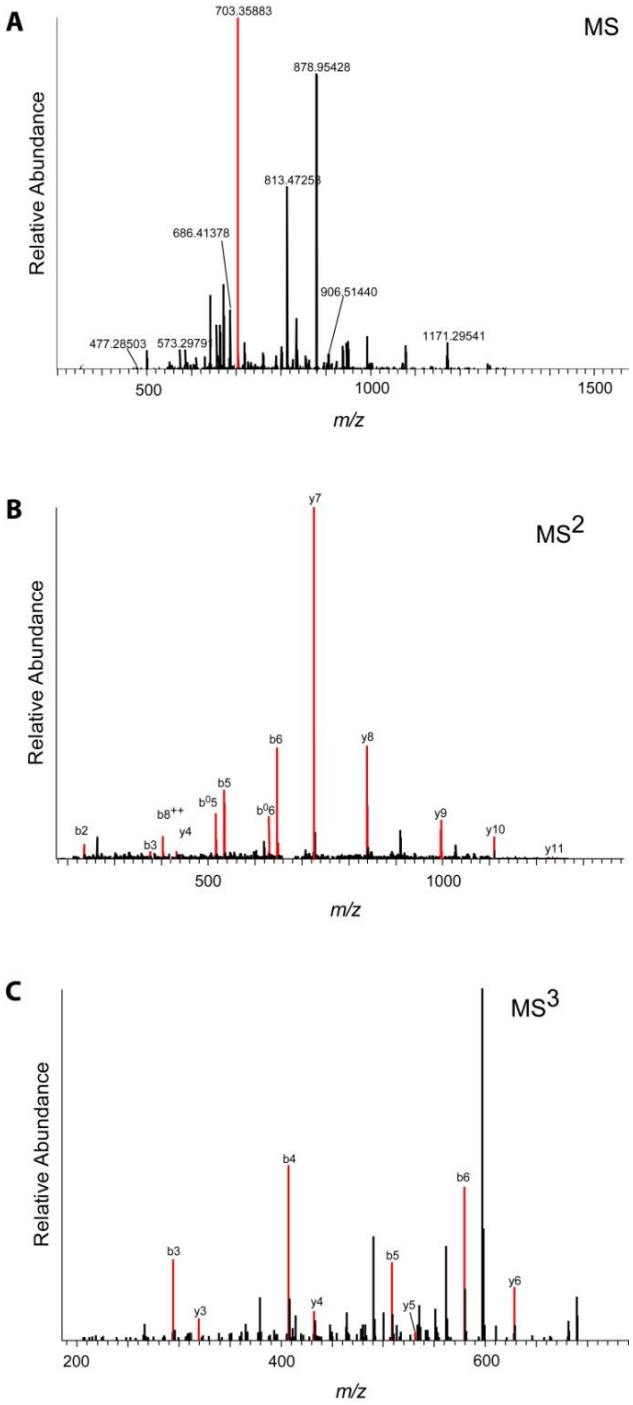


Figure 6.3 High confidence peptide identification by two consecutive stages of mass spectrometric fragmentation (MS^3). The precursor of a peptide, YVISAIPPVLTAK **(A)**, was selected for fragmentation from a full scan of mass to charge ratio range (shown in *red*). A fragment of the above, y_7 ion **(B)**, was subsequently fragmented. A characteristic pattern for charge-directed fragmentation is observed in the MS^3 spectra **(C)** and confirms the identification of the above peptide.

checked by EPICenter[237] and manually verified. EPICenter organizes all entries with shared peptides into a single group (protein group). The protein that contains most of the peptides is selected as an anchor, and all group members that are identified by at least one distinct and separate peptide are marked as conclusively identified. We counted proteins as *identified* only when a protein had at least one distinct peptide. If a protein group consisted of only isoforms or overlapping database entries indistinguishable by MS, then only the anchor protein was counted, thus the number of identified proteins is a lower boundary of the actual value.

From 45 LC-MS/MS/MS runs, 182,271 MS/MS spectra were submitted to Mascot database search and 52,585 MS/MS spectra satisfied the criteria for peptide validation. Among these, 22,706 represent unique sequences in the four cellular fractions. A total of 3,287 proteins were identified from four fractions with 8,953 unique peptides. Of all proteins, 20.4% were identified with single peptide identification and two stages of peptide fragmentation. Interestingly, 71.3% of the total was identified within only one cellular fraction whereas 16.7%, 7.8% and 4.2% were identified within 2, 3 and 4 fractions, respectively. This confirms the relatively high quality of organellar separation as already suggested by marker analysis in the western blotting experiments in Fig. 2. In two previous studies analyzing several mouse tissues each similarly separated into four subcellular compartments, protein overlap among compartments was deeper and fewer than 50% of total identified proteins were specific to one compartment[243,247].

6.3.2 Depth and Coverage of the 3T3-L1 Adipocyte Proteome assessed by Comprehensive Bioinformatics

6.3.2.1 Qualitative comparison with earlier studies

We compared our proteome dataset with the recently published mouse liver organelle proteome map[242] and the above mentioned study of six mouse tissues (brain, heart, kidney, liver, lung and placenta)[243]. As shown in Figure 6.4A, more than two third of the proteins identified by us in adipocytes overlapped with these other proteome. These proteins are candidates for the 'household proteome', i.e. proteins performing general cellular functions and therefore present in different cell lines and tissues. However, the proportion of proteins specific to adipocytes in our study (28.7%) is also relatively high. In contrast, in a previous study that compared six human

cell lines, specific proteins (proteins that were exclusively found in a single cell line) account for only 6% to 36% of all identified proteins[248]. As shown in Figure 6.4B, nearly half of our cytoplasm proteins overlapped with the combined cytoplasmic proteins from six human cell lines[248]. Secreted proteins were enriched approximately 3.5 fold in identified cytoplasmic proteins from adipocyte compared with cytoplasmic proteins from the six cell line proteome.

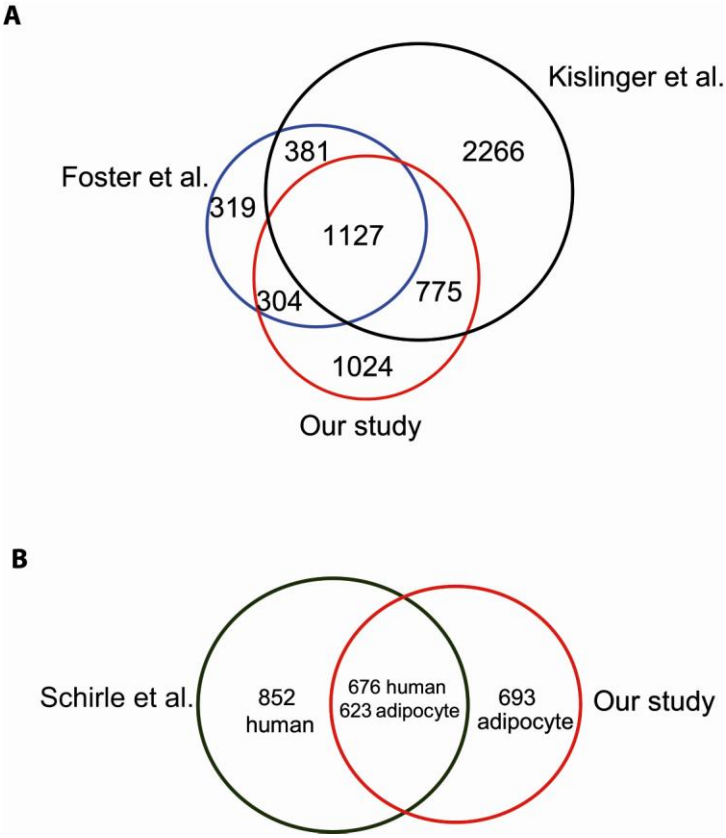


Figure 6.4 Two-thirds of adipocyte cell line proteins are also found in recent mouse organelle studies. **(A)** Mouse proteins reported recently in membrane-enclosed organelles of mouse liver (Foster *et al.*[242]) and a study of six mouse tissues (but without fat tissue) (Kislinger *et al.*[243]) were BLASTed against identified proteins in the current study by ProteinCenter (Proxeon Bioinformatics). Only proteins with at least 95% identity were considered to match. **(B)** The adipocyte cytoplasmic proteins were "BLASTed" against the combined cytoplasmic six-cell line proteome (Schirle *et al.*[248]). Only proteins with at least 80% sequence identity were considered to match.

Figure 6.4 clearly shows our proteome dataset contains many proteins which were not identified in previous large-scale proteome analysis using both tissues and cell lines. Our result may reflect

the depth of high-confidence analysis now possible and/or the specificity of adipocyte for their critical role in energy balance and whole body homeostasis.

6.3.2.2 Microarray comparison precludes any abundance related bias in proteome identification

Previously microarray studies have been undertaken to unravel various aspects of 3T3-L1 adipocyte differentiation, development and function[249-252] and they serve as a useful resource for gaining insights into mRNA expression and cellular dynamics. Moreover microarray studies provide an estimate of the transcriptome in a particular biological state at any given time, and we wished to use this data as a reference for estimating the coverage and depth of our large-scale proteome study. We analyzed the gene expression levels of normal 3T3-L1 adipocyte from Affymetrix microarray data generated in the Diabetes Genome Anatomy Project (DGAP). The available dataset was in triplicates for each of the MGU_74A, B, C Affymetrix array types. We combined the three array type datasets for our analysis. In total they contain 37,886 probe sets of which 7,656 were deemed 'present' by using the 66% Present (P) call criterion (see section 5.2.8). Out of these 7,656 probes, 2182 could be mapped to our identified proteome. We then divided the genes judged to be expressed in 3T3-L1 according to the microarray data into two groups: those identified in our study and those not identified (Figure 6.5A). If proteomics was biased to detect only high abundance proteins, we would expect a large difference in mRNA signal between the two groups. Remarkably, the distribution of mRNA expression levels was less than 2-fold higher for the genes whose products were detected in our proteomic analysis as compared to those that were not identified. To further substantiate this finding, we also compared our proteome data with the Gene Atlas V2 mouse microarray data for adipose tissue[253]. Again we observed that the distribution of mRNA expression level was less than 2-fold higher for the genes whose products were detected in our proteomic analysis as compared to those that were not (Figure 6.5B). This suggests that proteomics experiments, despite remaining limitations in complex mixture analysis[254], have become quite comprehensive and able to detect low-abundance proteins in cellular proteomes. Our previous study on mouse tissue mitochondria, in contrast, still showed a substantial tendency of mass spectrometry to preferentially detect products of high abundance messages[57].

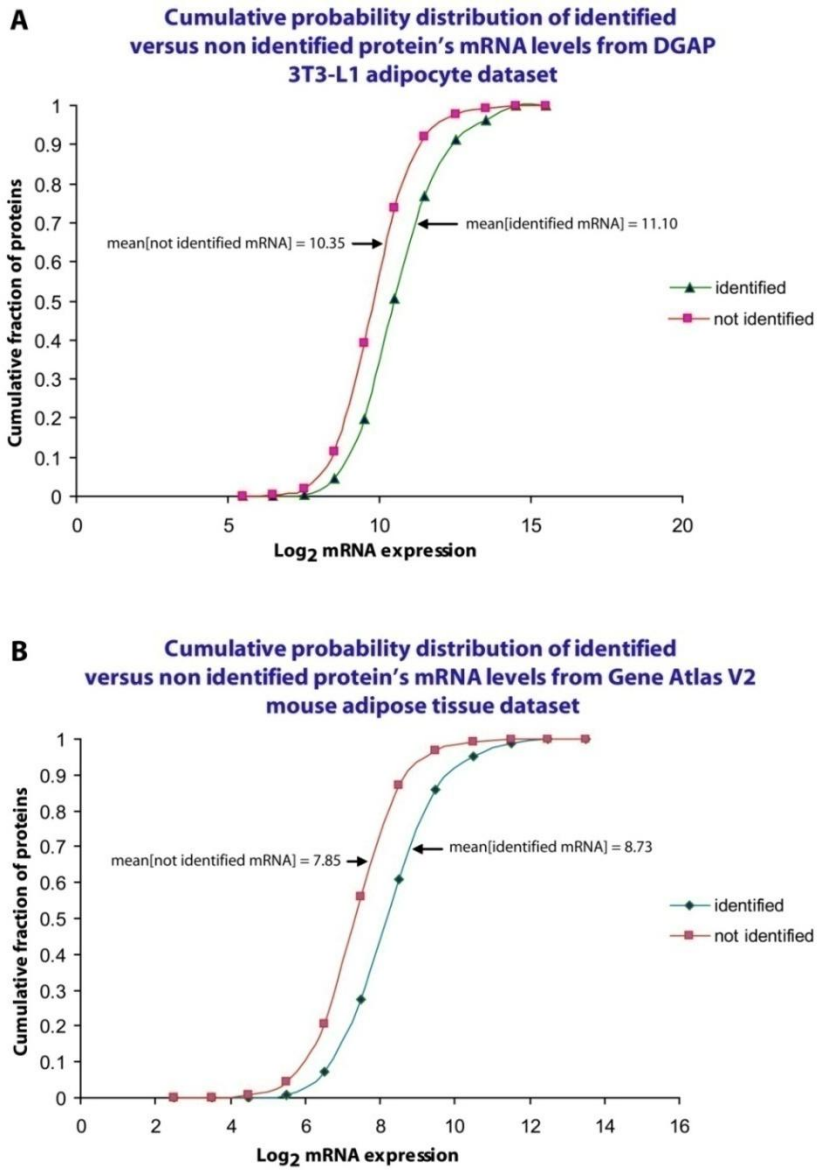


Figure 6.5 Cumulative probability distribution of the mRNA levels of identified *versus* not identified proteins. (A) Proteins identified in adipocytes were mapped onto DGAP 3T3-L1 mRNA data. The cumulative probability distributions of mRNA abundance for the genes whose protein products were detected (*green*) or not detected (*pink*) by proteomics are shown. The mean expression levels for both groups are indicated. **(B)** Proteins identified in adipocytes were mapped to GNF mouse atlas V2. The cumulative probability distributions of mRNA abundance for the genes whose protein products were detected (*green*) or not detected (*red*) by proteomics are shown.

6.3.2.3 Coverage of proteome in terms of pathways and annotated complexes

We then analyzed the coverage of our proteome in terms of protein complex and pathways known to be present in adipocytes. We mapped identified proteins onto KEGG pathways[255]. As shown in Figure 6.6A and 6.6B, almost all of the known proteins in the ribosome and proteasome complexes were identified. It is interesting that in the membrane fraction only the core 20S proteasome and not the 19S (PA700) and 11S (Psme1-3) regulatory complex were identified (Figure 6.6A).

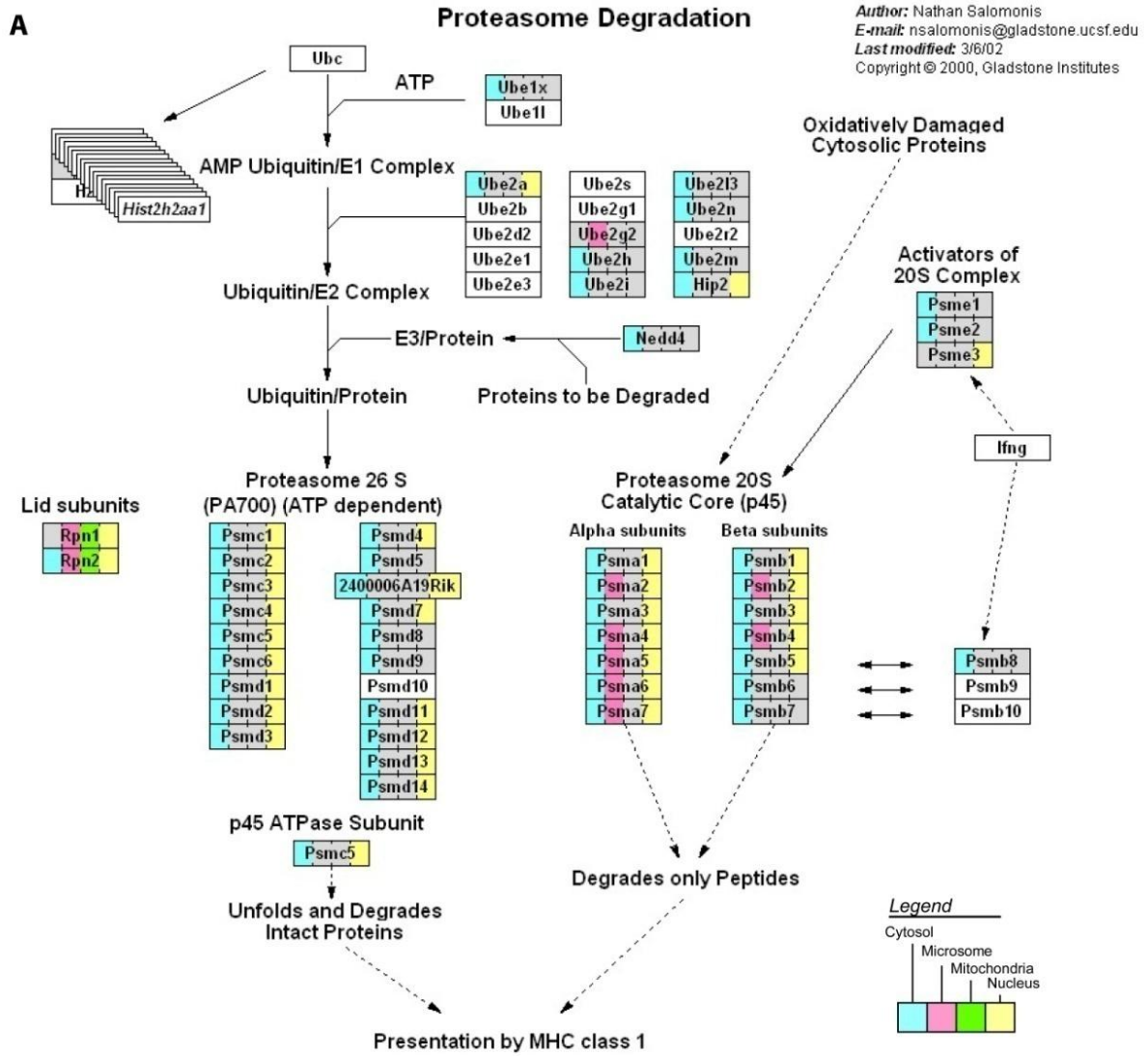


Figure 6.6 Pathway mapping of identified proteins in the four subcellular fractions. Proteins identified in four cellular compartments (cytosol, membrane, mitochondria, and nuclei) were mapped to known pathways using GenMAPP. **(A)** Proteasome degradation; **(B)** Cytoplasmic ribosomal proteins; **(C)** Insulin signaling pathway. The compartments are represented by four different colors for each gene symbol.

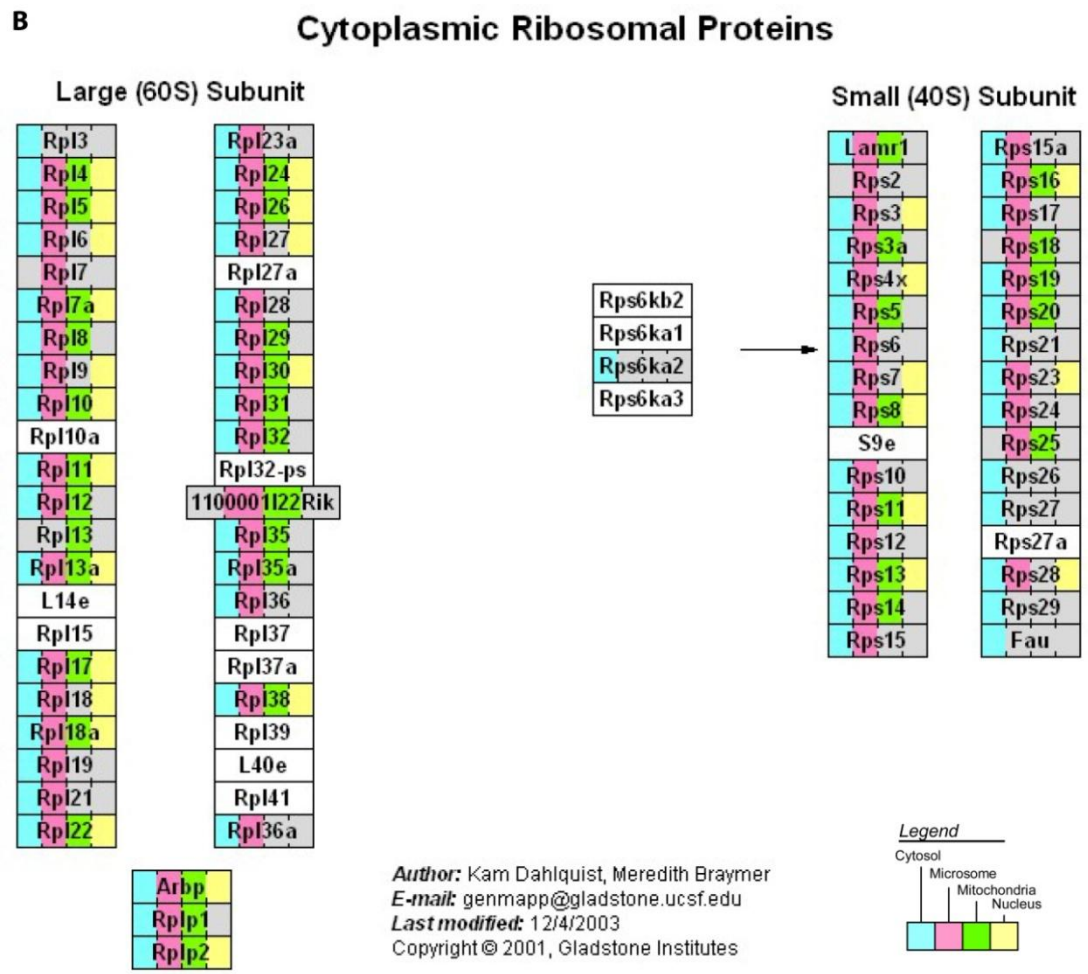


Figure 6.6(B) ***Continued***

The precise distribution of 20S and 19S complex at the cellular organelle level has not been reported in the literature. However, in agreement with our observations, Brooks *et al* reported 20S, 11S and 19S complexes localized predominantly in the cytosol and also localized in nuclear and membrane fractions prepared from rat liver[256]. The function of the 20S proteasome at the adipocyte membrane is unknown and would be interesting to elucidate in future studies. In contrast, less than half of the known proteins in the insulin pathway map were identified (Figure 6.6C). Coverage of kinases and transcription factors was low, while we detected more than half of the proteins related to vesicular trafficking. Interestingly, in the analysis of 3T3-L1 microarray data half of the key components of the insulin signaling pathway were also flagged as not expressed (i.e. filtered out by microarray quality measures when using the Present P calls > 66%

criterion). This suggests that the limitation of detection of low abundant proteins/genes such as kinases and transcription factors are no worse for MS based proteomics then for other high-throughput "omics" technologies.

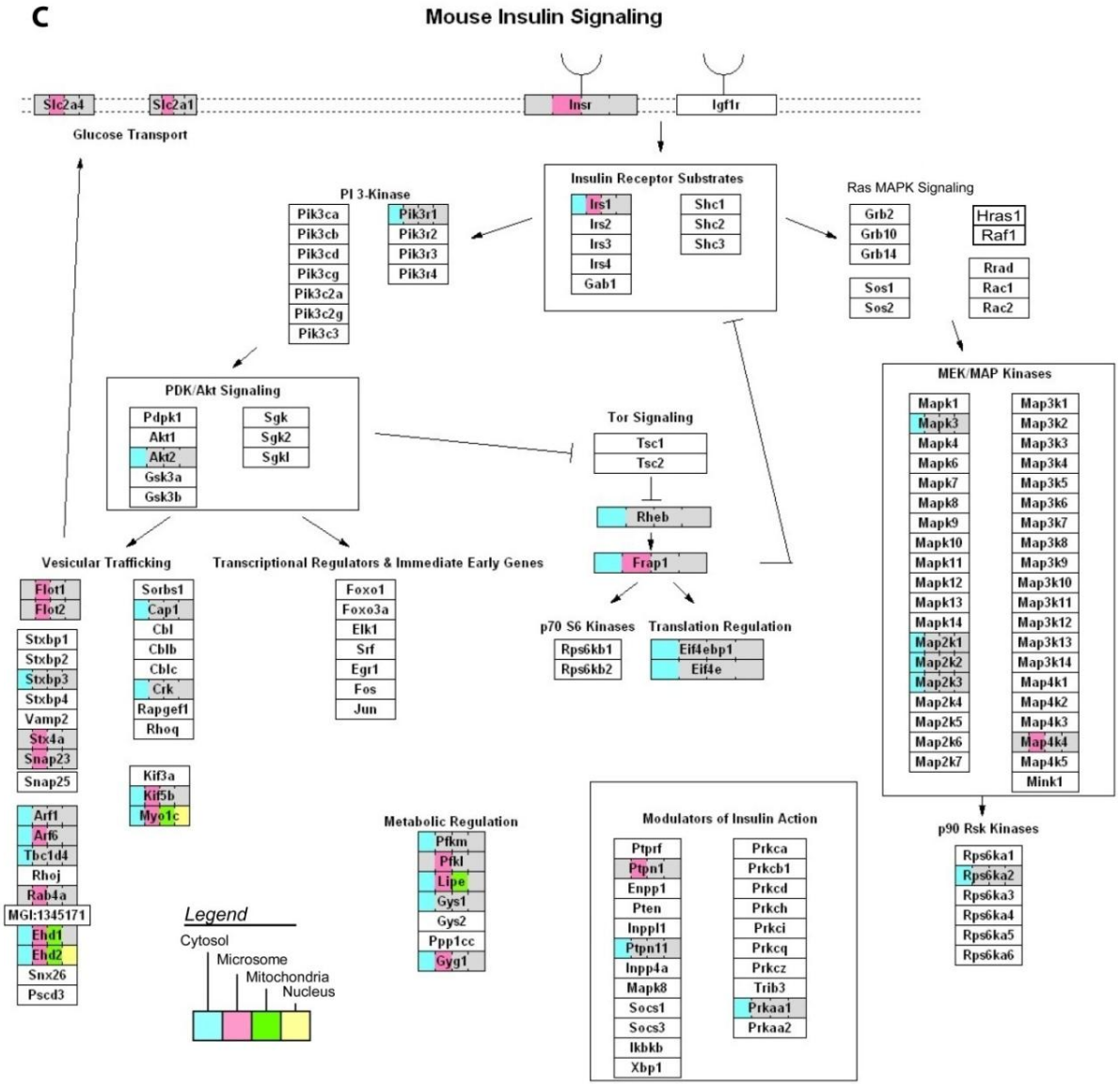


Figure 6.6(C) ***Continued***

6.3.3 Visual interpretation of proteome sub-cellular localization by hierarchical clustering and its concordance with earlier studies and genome wide annotations

In order to compare our subcellular fractions and detection coverage we benchmarked our data against previously reported large-scale proteomic analyses of mouse organelles[242,243] and Gene Ontology annotations. For our proteome, we built a cellular compartment distribution matrix (3,287 proteins by 4 fractions) by first counting peptides of each protein in each fraction. Then we normalized the data to arrive at a probability matrix for the distribution of each protein in the four compartments (see Section 5.2.11). Hierarchical clustering of this matrix shows that more than 70% of the proteins localized to four clusters with propensity for one specific fraction (Figure 6.7). A fifth cluster contained proteins for which there was no clear pattern of distribution. Additionally, we overlaid the data of the two large-scale mouse tissue experiments as well as cellular compartment annotations from Gene Ontology on the clustered dendrogram. As seen in the figure, most of the adipocyte cytosolic proteome showed high concordance with the experimental studies. Similarly, the membrane cluster shows good agreement. For the miotochondral and nuclear fractions there is excellent correlation, however the depth of our study was much greater. (Only the top part of our clusters detected counterparts in the other studies.) The GO annotations agreed well for the mitochondrial and nuclear specific clusters. No major enrichment is seen for the membrane and cytoplasmic fractions, which seem to be less well annotated in GO.

6.3.4 Protein Domain Enrichment for Insights into Protein Function

Classification of proteins based on their amino acid sequence or three-dimensional structure is one of the most established practices in protein science and also adopted by current large-scale structural genomics endeavors[257]. Moreover knowledge of independently folding protein domains can provide useful pointers into the complex interplay of proteome interactions and regulation by post translational modifications[258]. In order to obtain an additional perspective of the adipocyte proteome we performed InterPro domain enrichment analysis using our adipocyte proteome data set and the proteome data sets obtained from six mouse tissues[243] and extracted InterPro domains enriched only in the adipocyte proteome (Figure 6.8).

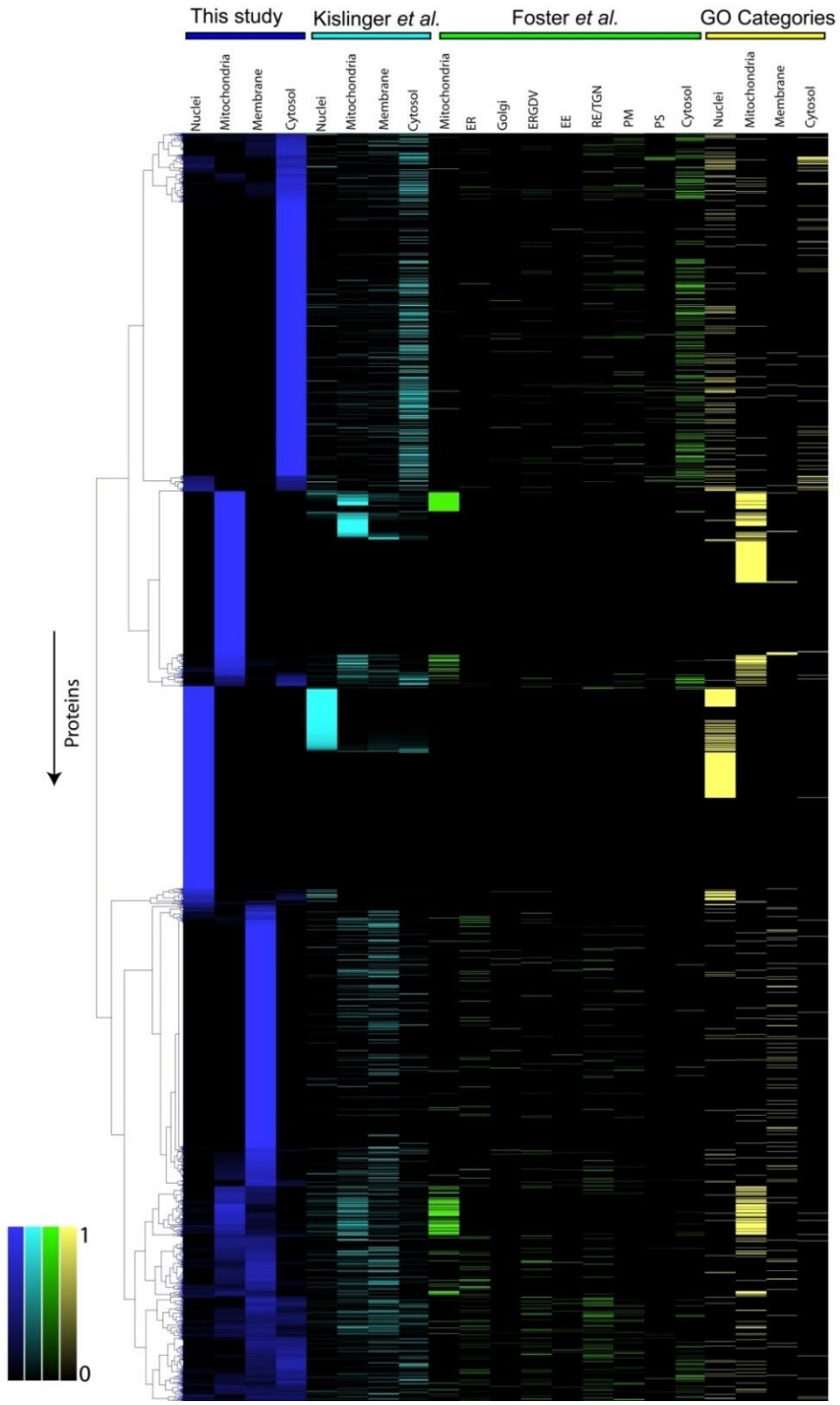


Figure 6.7 Concordance among subcellular location of our study, recently published mouse organelle datasets, and Gene Ontology annotation. One-dimensional hierarchical cluster dendrogram for the 3T3-L1 adipocyte cellular compartment profiles overlaid with data from recently reported large scale proteomics studies and GO cellular compartment terms is shown. The *dark blue color* represents the 3T3-L1 adipocyte proteome; *light blue* corresponds to the mouse tissue proteome study data by Kislinger *et al.*[243]; *light green* corresponds to the liver organelle protein study data by Foster *et al.*[58], *yellow* corresponds to the gene ontology (GO) cellular compartment annotations for the four compartments (nuclei, mitochondria, membrane, and cytosol) in our study

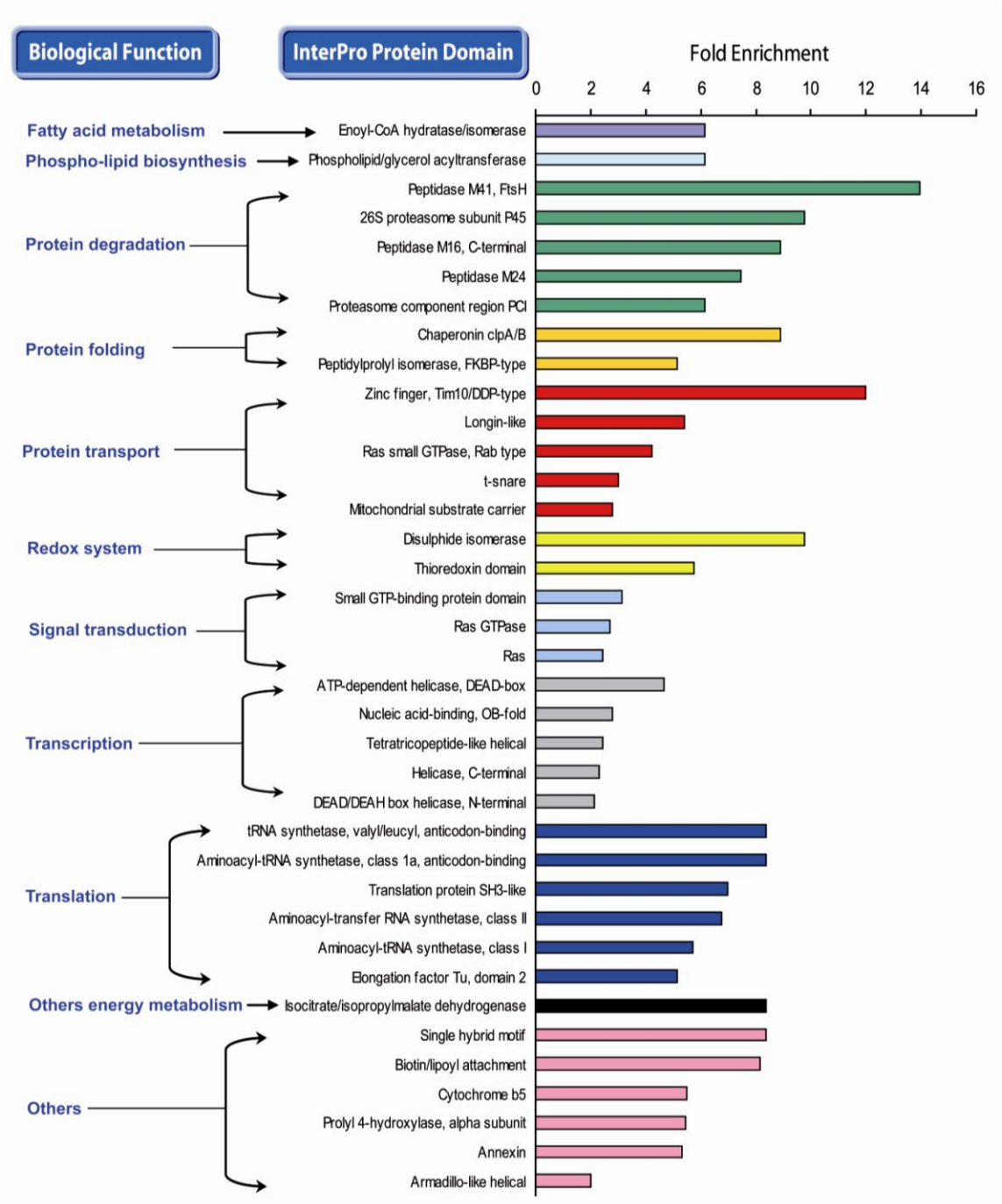


Figure 6.8 Significantly over-represented InterPro domains for the set of identified adipocyte proteins. Significantly over-represented InterPro terms with $p < 0.001$ and not also over-represented in the mouse tissue analysis are shown. For each InterPro term the bar shows the enrichment -fold ratio for the identified 3T3-L1 proteome in our survey with respect to the InterPro annotations of the entire mouse proteome. The InterPro terms are further grouped by representative biological function shown with text in *blue*. *SH3*, Src homology 3; *FKBP*, FK506-binding protein. *OB*, oligosaccharide/oligonucleotide-binding

The function of these enriched domains were mainly related to signal transduction, redox system, protein transport, translation, transcription, protein degradation, fatty acid metabolism, phospho lipid biosynthesis, in agreement with the enriched GO terms described before. Enrichment of redox related domains, such as the thioredoxin domain is interesting, because it has been suggested that the reduced redox state encourages triglyceride synthesis, adipocyte differentiation, and the development of adipose tissue[259] while an increase in the markers of systemic oxidative stress has been associated with obesity and metabolic syndrome[260]. Similarly the domains related to vesicular protein transport, such as Ras small GTPase, Rab type, t-snare, Longin-like and a domain Zinc finger, Tim10/DDP-type which is related to protein import into mitochondrial inner membrane were substantially enriched.

Domains related to transcription and translation were also enriched, specially three domains of aminoacyl-transfer RNA synthetases, namely Aminoacyl-tRNA synthetase, class 1a, anticodon-binding, Aminoacyl-transfer RNA synthetase, class II, and Aminoacyl-tRNA synthetase, class I. Transcription and translation are basic functions of the cell, thus proteins related to such functions are generally thought to be housekeeping proteins. This observed enrichment may not reflect basic adipocyte biology but simply the fact that a rapidly growing cell line needs to express more proteins than the comparatively more inert tissue. Further insights may be obtained by quantitative study of protein expression in different cell lines and tissues, and by creating a protein expression atlas similar to a gene expression atlas[261,262].

6.3.5 An integrative genomics approach for protein prioritization analysis of vesicular trafficking in adipocytes

One of the important features of adipocytes is insulin regulated glucose uptake. In adipocytes, the majority of this glucose uptake results from the translocation of the glucose transporter 4 (GLUT4) to the cell surface membrane. Since the cloning of GLUT4 in 1989 in several laboratories[263-267] numerous studies have attempted to elucidate the molecular basis of insulin receptor-signaling pathway and membrane-trafficking processes. One of the unresolved questions is the connection between Akt activation and GLUT4 translocation. GO term enrichment analysis revealed that protein transport was enriched in the adipocyte proteome (see above) and some of the identified vesicular trafficking proteins are known to be involved in the

insulin signaling pathway (Figure 6C). As 26%, 35% and 36% of identified proteins in our study were not annotated by Gene Ontology molecular function, biological process and cellular component categories, respectively, we tried to predict candidate proteins involved in GLUT4 translocation using a bioinformatics approach. Very recently, an algorithm termed Endeavour was developed for gene prioritization to rank genes involved in human diseases and biological processes[8]. The concept of prioritization by Endeavour is that candidate test genes are ranked based on their similarity with a set of known training genes. The similarity measure is in turn calculated by integrating functional, process, gene ontology (GO), pathway and sequence similarity information obtained from diverse data sources. As training genes we choose 29 genes involved in vesicular trafficking which are on the map of Figure 6C. We used 2,990 proteins which were identified and mapped to human Ensembl gene Identifiers in our study as test genes. For 41 gene products we obtained highly significant values ($p < 0.0002$) for association with our set of proteins known to be involved in vesicular traffic (Table 6.1). Candidate proteins highly ranked by Endeavour contain many ras-related GTP-binding proteins (Rabs) and soluble N-ethylmaleimide-sensitive factor attachment protein receptors (SNAREs). While it is not surprising that these proteins are involved in vesicular trafficking, they do serve as a positive control of the algorithm. We found that proteins recently associated with insulin signaling or GLUT4 translocation were ranked high in the candidate proteins. For example, Rab10, Rab14, Rab2, vesicle transport through interaction with t-SNAREs 1B homolog, vacuolar protein sorting 45, vesicle-associated membrane protein 8 and syntaxin 12 are known to be contained in GLUT4 vesicles[268,269]. Rab2, Rab10 and Rab14 were identified as targets of Akt substrate of 160-kDa (AS160) whereas Rab4 was reported to be involved in insulin-induced GLUT4 translocation[270]. ADP-ribosylation factor 5 (Arf5) was observed to exhibit modest re-distribution to the plasma membrane in response to insulin[271] and cdc42, a Rho GTPase family member mediates insulin signaling to glucose transport in 3T3-L1 adipocytes[272]. The above examples show that the protein prioritization by Endeavour is reasonable. By extension, candidates in Table 6.1 with no obvious connection to insulin signaling and GLUT4 are now excellent candidates for further functional study in this context.

IPI	Description	Gene symbol	p-value
IPI00116770	RAB18, member RAS oncogene family	Rab18	2.57E-06
IPI00221615	ADP-Ribosylation factor 5	Arf5	3.82E-06
IPI00271059	RAB4B, member RAS oncogene family	Rab4b	4.77E-06
IPI00132276	vesicle-associated membrane protein 3	Vamp3	6.03E-06
IPI00113849	cell division cycle 42 homolog	Cdc42	6.83E-06
IPI00331663	unnamed protein product	Arf4	7.24E-06
IPI00122965	RAB3A, member RAS oncogene family	Rab3a	8.52E-06
IPI00114560	RAB1, member RAS oncogene family	Rab1	9.91E-06
IPI00137227	RAB2, member RAS oncogene family	Rab2	1.07E-05
IPI00132410	RAB5A, member RAS oncogene family	Rab5a	1.13E-05
IPI00118217	Syntaxin-7.	Stx7	1.24E-05
IPI00453589	Vesicle-associated membrane protein 8 (VAMP-8)	Vamp8	1.43E-05
IPI00126042	RAB14, member RAS oncogene family	Rab14	1.49E-05
IPI00230011	Rab6 protein	Rab6	1.61E-05
IPI00137647	synaptobrevin like 1	Sybl1	1.71E-05
IPI00321581	GS32 protein	Snap29	1.90E-05
IPI00124291	vacuolar protein sorting 45	Vps45	2.35E-05
IPI00331128	cell line NK14 derived transforming oncogene	Rab8a	2.36E-05
IPI00116729	Ras-related protein Rab-22A (Rab-22) (Rab-14)	Rab22a	3.24E-05
IPI00125880	protein kinase C and casein kinase substrate in neurons 2	Pacsin2	4.34E-05
IPI00111416	syntaxin 12	Stx12	4.87E-05
IPI00469799	splice isoform 2 of golgi autoantigen, golgin subfamily A member 3	Golga3	4.91E-05
IPI00416303	aminopeptidase-like 1	Npepl1	5.52E-05
IPI00224219	sec1 family domain containing 1	Scfd1	6.32E-05
IPI00109506	unnamed protein product	Stx6	6.79E-05
IPI00224518	Ras-related protein Rab-5C	Rab5c	7.15E-05
IPI00225581	Dedicator of cytokinesis protein 1(Fragment)	Dock1	7.19E-05
IPI00134941	c-K-ras2 protein	Kras	9.95E-05
IPI00116558	ras homolog gene family, member G	Rhog	9.96E-05
IPI00121335	thymoma viral proto-oncogene 2	Akt2	1.09E-04
IPI00116688	RAB3D, member RAS oncogene family	Rab3d	1.09E-04
IPI00378015	Drebrin-like protein (SH3 domain-containing protein 7)	Dbnl	1.17E-04
IPI00130118	RAB10, member RAS oncogene family	Rab10	1.28E-04
IPI00378145	RAB6B, member RAS oncogene family	Rab6b	1.30E-04
IPI00453776	early endosome antigen 1	Eea1	1.40E-04
IPI00453771	prenylated SNARE protein Ykt6	Ykt6	1.41E-04
IPI00132685	blocked early in transport 1 homolog	Bet1	1.41E-04
IPI00229483	SEC24 related gene family, member C	Sec24c	1.43E-04
IPI00131445	Dynamin-2 (Dynamin UDNM)	Dnm2	1.49E-04
IPI00331284	vesicle transport through interaction with t-SNAREs 1B homolog	Vti1b	1.50E-04
IPI00130554	SNAP-associated protein	Snapap	1.94E-04

TABLE 6.1 Putative proteins involved in the vesicular trafficking in insulin signaling predicted by the method of gene prioritization

6.4 Discussion

Adipocytes are central players in energy metabolism and the obesity epidemic, yet their protein composition remains largely unexplored. Elucidating the protein composition of this versatile cell type is the first step towards understanding its role in various cellular processes and disease pathophysiologies. Our adipocyte proteomics study using enriched cellular compartments and state of the art mass spectrometry, involving very high mass accuracy and two stages of mass spectrometric fragmentation, allowed us to establish a high-confidence set of adipocyte proteins consisting of 3,287 proteins. Our analysis provides one of the largest and most confident set of proteins present in any cell line or tissue. Not only the identified protein list, but also the data on putative proteins - involved in vesicular trafficking in insulin signaling reported here should serve as a useful reference for more extensive experimental characterization of adipocyte functions. In order to share the data presented in this study, we have made the adipocyte proteome accessible at the Max-Planck Unified Proteome database (MAPU database, http://proteome.biochem.mpg.de/adipo/)[273]. While the MS technologies are already in place to elucidate comprehensive proteomes of model organisms[64], continuing advances in the sensitivity and automation of MS-based proteomics will soon make acquisition of complex cellular proteomics routine. The analytical and bioinformatics analysis framework applied here can then serve as the template for processing and data mining of such cellular proteomes.

7. Comparative proteomic phenotyping to assess functional differences between primary hepatocyte and the Hepa1-6 cell line

This work is included in a manuscript accepted for publication in *Molecular Cellular Proteomics*:

Cuiping Pan[φ], **Chanchal Kumar**[φ], Sebastian Bohl, Ursula Klingmueller, Matthias Mann

Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type specific functions

[φ] These authors contributed equally to this work

7.1 Introduction

The development of tissue culture techniques and establishment of cell lines has been indispensable for biological research for several decades[274]. However, disadvantages of cell lines are that they are usually derived from tumors and have adapted to growth in culture. Although cell culture tries to create a close-to-physiology milieu by adding appropriate amounts of salt, glucose, amino acids, vitamins, and serum, the lack of tissue architecture and heterogeneous population of cell types often abolishes cell-cell interaction, secretion, and other functions based on tissue context. Cells in culture are prone to genotypic and phenotypic drifting. Thereby cell lines can lose tissue specific functions and acquire a molecular phenotype quite different from cells *in vivo*. Acceptance of cell lines as model for biological function varies between fields. Cell biological studies on basic mechanisms, such as the cell cycle are routinely and overwhelmingly carried out in long-established cell lines[275]. This is particularly the case for microscopy studies, including large-scale siRNA screens with imaging read out. In contrast, there is substantial controversy of how well cell lines – which are often established from late stage cancer – preserve aspects of the disease and whether or not they should be used in cancer drug development[276-278]. Thus animal experiments or studies in primary cell lines are often preferred despite their added complexity. Accurate molecular phenotypes to determine if the function to be investigated is preserved in cell lines would enable a rational choice of the most appropriate experimental

system[279]. In biotechnology and the pharmaceutical industry this goal obtains added urgency in light of efforts to reduce animal experimentation to a minimum.

In this work we ask how primary cells and cell lines differ in their functions. This question has been addressed by comparing gene expression profiles at the transcriptome level in a substantial body of literature (for recent examples see[280,281]). However, transcriptome studies are not quantitative with respect to changes at the proteome level. Ideally, the different molecular phenotypes should be assessed by quantitatively comparing the proteomes of the primary cells vs. the cell lines. Here we report such a study and develop an algorithm to extract functional phenotypes from the resulting differential protein distributions.

7.2 Material and Methods

7.2.1 Materials and reagents

Mouse hepatoma cell line Hepa1-6 was obtained from American Type Culture Collection (ATCC). L-arginine, L-lysine, L-$^{13}C_6{}^{15}N_4$-arginine and L-$^{13}C_6{}^{15}N_2$-lysine were purchased from Sigma-Aldrich. Chemicals for the ‘in solution’ and ‘in gel’ digests were purchased from Sigma-Aldrich, Endoproteinase Lys-C was obtained from Waco and sequencing grade modified trypsin was from Promega.

7.2.2 Isolation of mouse primary hepatocytes

Isolation and culture of mouse hepatocytes was performed according to standard operation procedures[282]. For biological and analytical reproducibility, primary hepatocytes were isolated from two mice and processed separately. After cultivation for 14 hours, the cells were placed on ice and the medium was removed. The cells were lysed in RIPA buffer.

7.2.3 SILAC labeling of mouse hepatoma cell line Hepa1-6

Hepa1-6 cells were grown in SILAC “light” (L-arginine and L-lysine) and “heavy” (L-$^{13}C_6{}^{15}N_4$-arginine and L-$^{13}C_6{}^{15}N_2$-lysine) conditions for 8 passages before the experiment. This period lasted around 3 weeks. Unless stated otherwise, cell culture medium contained 4.5 g/L glucose

by following the standard culture condition from ATCC. Other cell culture conditions were essentially the same as described[35].

7.2.4 Fluorescence microscopy

Primary hepatocytes were isolated and seeded at a density of 2×10^5 cells per well in collagen I coated 12-well plates. Hepa1-6 cells were grown to a density of 2×10^5 cells per well in 12-well plates. Cells were stained for 15min with 15nM Mitotracker Orange CMTMRos (Invitrogen) and 1:5000 Hoechst 333342 (Sigma-Aldrich) at 37℃ in the corresponding cultivation medium. After three washing steps in cultivation medium cells were viewed with a ZeissAxioVert 200M fluorescence microscope using a 40× LD-Plan Neofluor objective (n.a. 1.5). Cells were viewed under visible light, or excited with 345nm (Hoechst 333342) or 550nm (Mitotracker).

7.2.5 Protein harvest, digestion

Primary hepatocytes and Hepa1-6 cells were lysed in a buffer containing 1% NP-40, 0.1% sodium deoxycholate, 150 mM NaCl, 1mM EDTA, 50 mM Tris, pH 7.5, 1 mM sodium orthavanadate, 5 mM NaF, 5 mM beta-glycerophosphate and protease inhibitors (Complete tablet, Roche Diagnostics). The lysates were centrifuged in cold with 17,000g for 15 minutes to pellet cellular debris. Supernatant was collected and a Bradford method was used to determine the protein concentrations. Equal amount of the proteins from the primary hepatocyte sample and Hepa1-6 sample were mixed, resulting in 100 µg proteins in total.

Protein mixtures were added with four volumes of methanol, one volume of chloroform and three volumes of distilled water in a sequential manner. The addition of each solvent was followed by a short vortex. After centrifugation of 20,000g for 1 minute, proteins were focused between organic and inorganic phases. The aqueous phase was discarded. Four starting volumes of methanol were added to the protein pellet followed by a short vortex. After spinning at 20,000g for 2 minutes, methanol was removed and the protein pellet was air-dried.

Precipitated proteins were redissolved in a buffer containing 6 M urea, 2 M thiourea, 10 mM Hepes, pH 7.5. Proteins were reduced with 1 mM dithiothreitol for 1 hour, alkylated with 5.5 mM iodoacetamide for 45 minutes in dark, and digested for four hours with endoproteinase Lys-C (1/50 w/w). After diluting four times with 20 mM ammonium bicarbonate, samples were

digested overnight with sequencing grade modified trypsin (1/50 w/w). The digestion was quenched by adding trifluoroacetic acid to reach pH <3.

7.2.6 Peptide preparation for mass spectrometry

Peptides were separated based on their isoelectric points in the Agilent 3100 OFFGEL Fractionator (Agilent, G3100AA) and the 3100 OFFGEL Low Res Kit, pH 3-10 (Agilent, 5188-6424) according to the manufacturer. Peptides were focused for 20 kVh at maximum current of 50A and maximum power of 200 mW. Each peptide fraction was mixed with 10 µl solvent containing 30% acetonitrile, 5% acetic acid and 10% trifluoroacetic acid. The resulting solution was loaded into C_{18} reverse phase StageTips[235]. Peptides were eluted from the StageTips by applying 80% acetonitrile, 0.5% acetic acid. Samples were dried down to 3 µl and mixed with equal volume of solvent containing 2% acetonitrile and 1% TFA. 5 µl samples were applied for LC-MS/MS analysis.

7.2.7 Mass spectrometry and data analysis

Samples were injected via autosampler into a 15-cm fused silica emitter (75-µm inner diameter; Proxeon Biosystems) packed in-house with reverse-phase ReproSil-Pur C18-AQ 3-µm resin[84] and eluted with nanoflow in Agilent 1200 liquid chromatography system (Agilent Technologies, Waldbronn, Germany). The gradient induced a linear increase of 4-40% acetonitrile in 0.5% acetic acid over 90 minutes. Eluted peptides were sprayed into a 7-T LTQ-FT or LTQ-Orbitrap mass spectrometer (Thermo Electron, Bremen, Germany) via a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark) and analyzed as described previously[84]. Raw MS spectra were processed using in-house developed software MaxQuant (version 1.0.7.4)[2,283] which performs peak list generation, SILAC- and extracted ion current-based quantitation, posterior error probability (PEP) and false discovery rate (FDR) based on search engine results, peptide to protein group assembly , and data filtration and presentation[283]. The derived peak list was searched with the Mascot search engine (version 2.1.04, Matrix Science, London, UK) against a concatenated database combining 52,326 proteins from International Protein Index (IPI) mouse protein database version 3.24, 27 commonly observed contaminants (forward database) and the reversed sequences of all proteins (reverse database). Carbamidomethylation was set as fixed

modification. Variable modifications included oxidation (M), N-acetylation (protein), pyro (N-term QC). Full tryptic specificity was required and up to three missed cleavages were allowed. Initial mass deviation of precursor ion and fragment ions were up to 10 ppm and 0.5 Da, respectively. Maximum peptide PEP was set to 0.1 and peptide FDR and protein FDR were set to 0.01.

7.2.8 Gene Ontology and KEGG enrichment analysis based hierarchical clustering

In the primary against cell line study the quantified proteome was divided into 5 quantiles corresponding to probability cutoffs of 0, 0.15, 0.25, 0.75, 0.85, and 1. The enrichment analysis for gene ontology (GO) biological process and cellular component were done separately for these quantiles with respect to the whole quantified proteome by conditional hypergeometric test available in the GOstats package[284] in the R statistical environment[285]. For hierarchical clustering we first collated all the categories obtained after enrichment along with their *p*-values, and then filtered for those categories which were at least enriched in one of the quantiles with *p*-value < 0.05. The categories which did not have a defined *p*-value after collation in any quantile because the reference category members were missing were provided a *p*-value of 1. This filtered *p*-value matrix was transformed by the function $x = -log_{10}$ (*p*-value). Finally these x values were transformed to *z*-score for each GO category by using the transformation $\frac{x-mean(x)}{sd(x)}$. These *z*-scores were then clustered by one-way hierarchical clustering using "Euclidean distance" as distance function and "Average Linkage clustering" method available in Genesis[241]. KEGG pathway enrichment analysis was done in the same way, except that the hypergeometric test was employed and the reference set was complete mouse KEGG annotation.

7.3 Results

7.3.1 Quantitative analysis of Hepa1-6 against primary hepatocytes

To characterize phenotypic differences between cell lines and primary cells, we SILAC-labeled[9,286] a murine hepatoma cell line, Hepa1-6[287], and compared its proteome to that of primary hepatocytes prepared according to standard operating procedures(SOP) established by the German systems biology competence network HepatoSys[282] (Figure 7.1). We used high

resolution MS to identify 3,400 proteins while quantifying more than 3,350 of them (see Section 7.2.7). We repeated the experiment with hepatocytes from another mouse and obtained excellent reproducibility (Pearson correlation coefficient 0.95; Figure 7.2). We then combined the two datasets and analyzed them together using stringent and unified criteria. At a false positive rate of less than one percent, a total of 4,063 proteins were identified and quantified between the two cell populations.

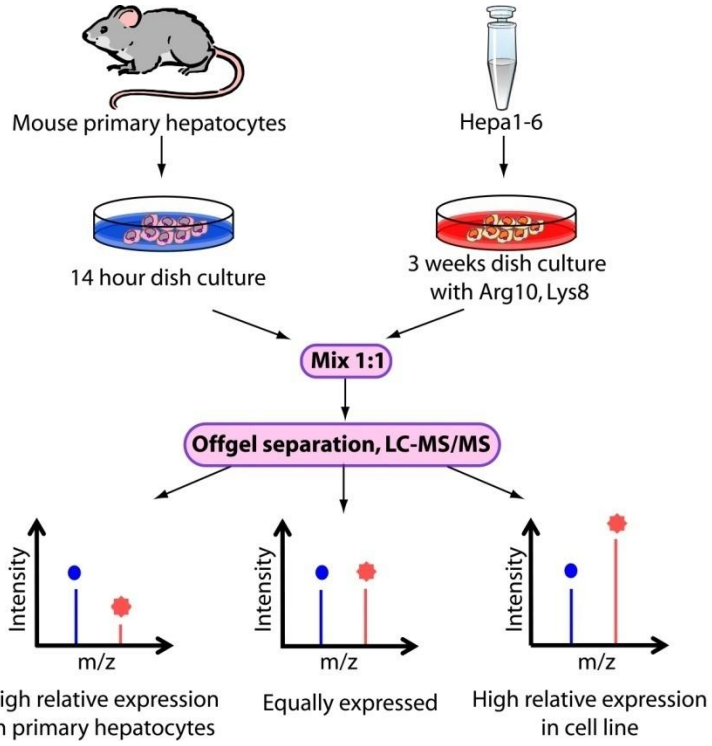


Figure 7.1 Strategy for comparing primary cells with immortalized cell lines. Primary hepatocytes were isolated and grown for 14 h. The Hepa1-6 cell line was completely SILAC-labeled with $^{13}C_6{}^{15}N_4$-arginine and $^{13}C_6{}^{15}N_2$-lysine. Cell extracts were combined and analyzed by online high-resolution MS on a linear ion trap Fourier transform instrument (LTQ-FT).

The primary and cell line proteomes overlap qualitatively but are very different quantitatively, with more than half of the proteome changing at least two-fold between the two conditions (Figure 7.3A,7.3B). Many proteins are expressed at much lower levels in the immortalized cell line than in the primary cells whereas comparatively few were up-regulated in Hepa1-6. This is surprising since cancer cells are thought to be de-differentiated and to express many genes

inappropriately. It is usually recommended to cultivate Hepa1-6 cells in high glucose medium (38 mM). Therefore we asked whether some of the observed phenotypic changes are attributable to this circumstance. To address this experimentally, we performed another SILAC experiment comparing cells cultured in high glucose against physiological glucose levels in mice (8 mM) for three weeks (Figure 7.4A).

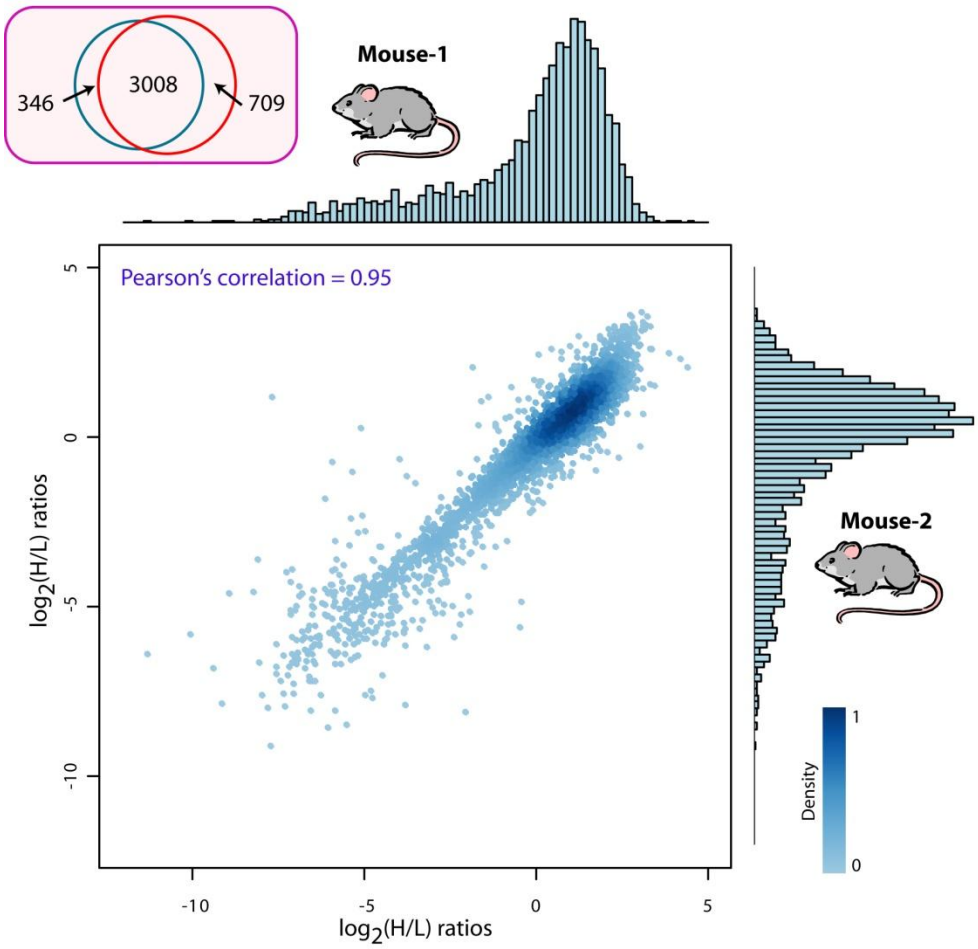


Figure 7.2 Replicate experiments of comparing Hepa1-6 cell line with primary hepatocytes from two mice achieved very high degree of reproducibility at both identification and quantitation level (about 4,000 proteins quantified; Person correlation coefficient 0.95).

In this experiment, there were hardly any overall changes in the proteome and 96% of the proteins were of constant abundance within a factor of two. This was also confirmed in a replicate experiment (Figure 7.4B). These results rule out a dominant role of the superphysiological glucose level in the proteome differences between primary cells and cell

lines. Furthermore, they demonstrate excellent quantitative accuracy of our experiment on a proteome-wide basis.

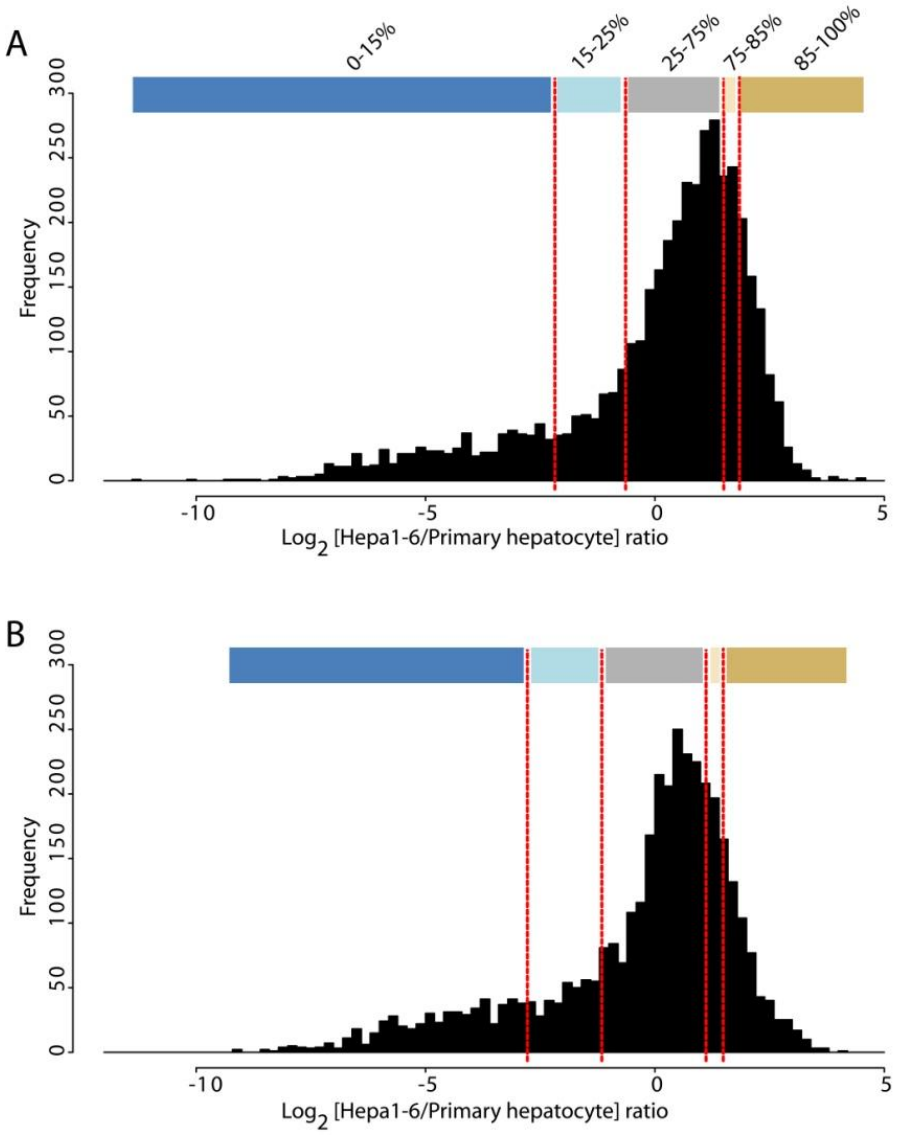


Figure 7.3 Fold-change distributions of the proteome (A) Quantitative comparison of the primary against the Hepa1-6 cell line proteome. The distribution was divided into five quantiles as follows. High relative expression in primary cells (0-15%, at least four-fold down-regulation), mostly expressed in primary cells (15-25%, -4 to -1.5 fold regulation), not highly regulated proteins (25-75%; - 1.5 to +2.8), mostly expressed in Hepa1-6 (75-85; 2.8 to 3.6 fold), highly expressed in Hepa1-6 (85-100%, more than 3.6 fold change). Color coding of these categories is indicated at the top of the panel. **(B)** Biological replicate of the experiment showing excellent reproducibility

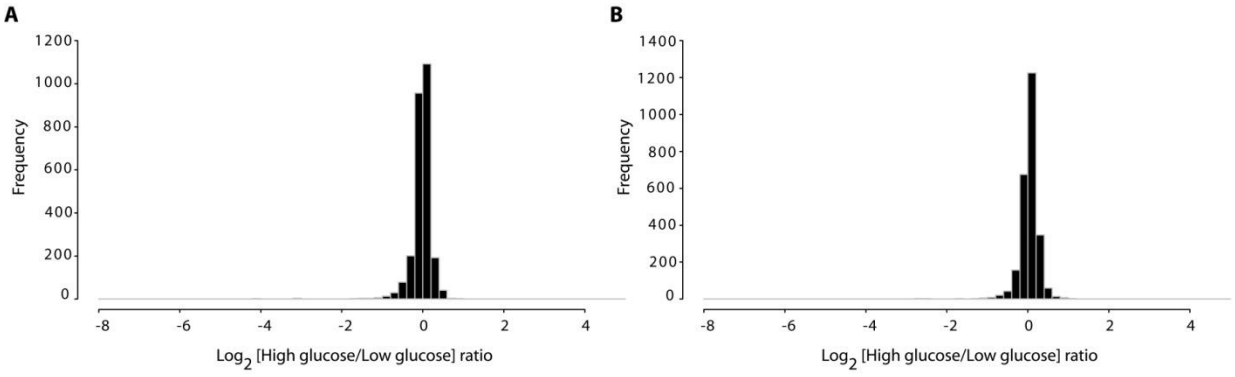


Figure 7.4 Quantitative comparison of the Hepa1-6 proteome cultured in high glucose (38 mM) and physiological glucose concentration (8 mM). Two independent comparisons **(A)** and **(B)** were performed starting from cell culture and SILAC labeling.

7.3.2 A novel bioinformatics method for proteomic phenotyping

To functionally understand the differences between the two cell populations, we divided the fold-change distribution between primary hepatocytes and the Hepa1-6 cell line into five quantiles according to relative protein expression (Figure. 7.3A, 7.3B). Each quantile was assessed separately for overrepresented pathways, biological processes and cellular components with Gene Ontology (GO) and KEGG pathway analysis[49,288] (Figure 7.5; Section 7.2.8). We retained each functional category that reached at least 95% statistical significance in one of the quantiles and then performed one-way unsupervised clustering of the p-values of the resulting categories (Fig. 7.6). This analysis differs from the more familiar clustering of overrepresented genes themselves, which is frequently employed in microarray-based experiments. It integrates the strength of statistical testing (taking p-values as input for clustering) with the intuitive simplicity of hierarchical clustering. By automatically classifying related processes and pathways based on their up or down-regulated protein measurements, it provides an unbiased global portrait of representative biological functions, enabling visual interpretation of the phenotype in terms of aggregate functional modules on a systems level.

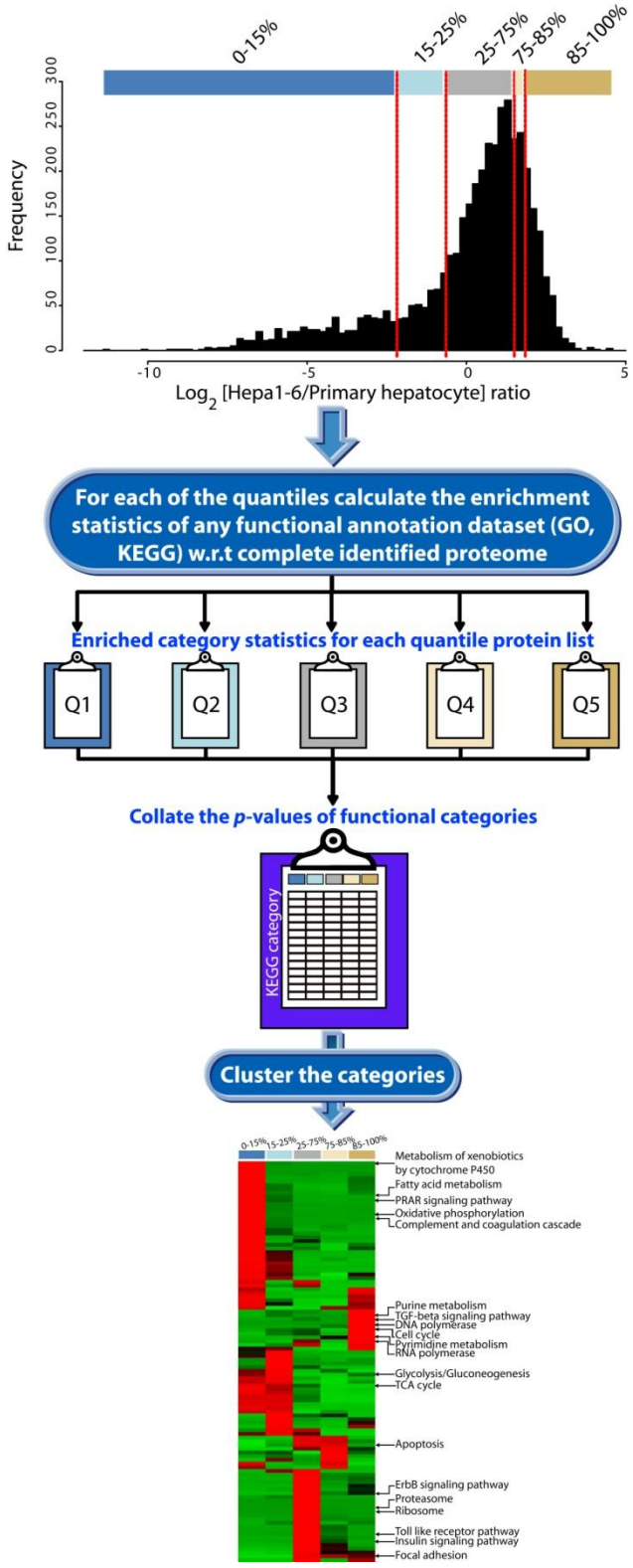


Figure 7.5 Bioinformatics workflow for proteomic phenotyping (detail in section 7.2.8)

We verified the robustness of these functional assignments by comparing the shared p-value matrix of the replicate experiments against each other. This correlation was 0.86 for KEGG, 0.85 for GO biological process and 0.92 for GO cellular compartment.

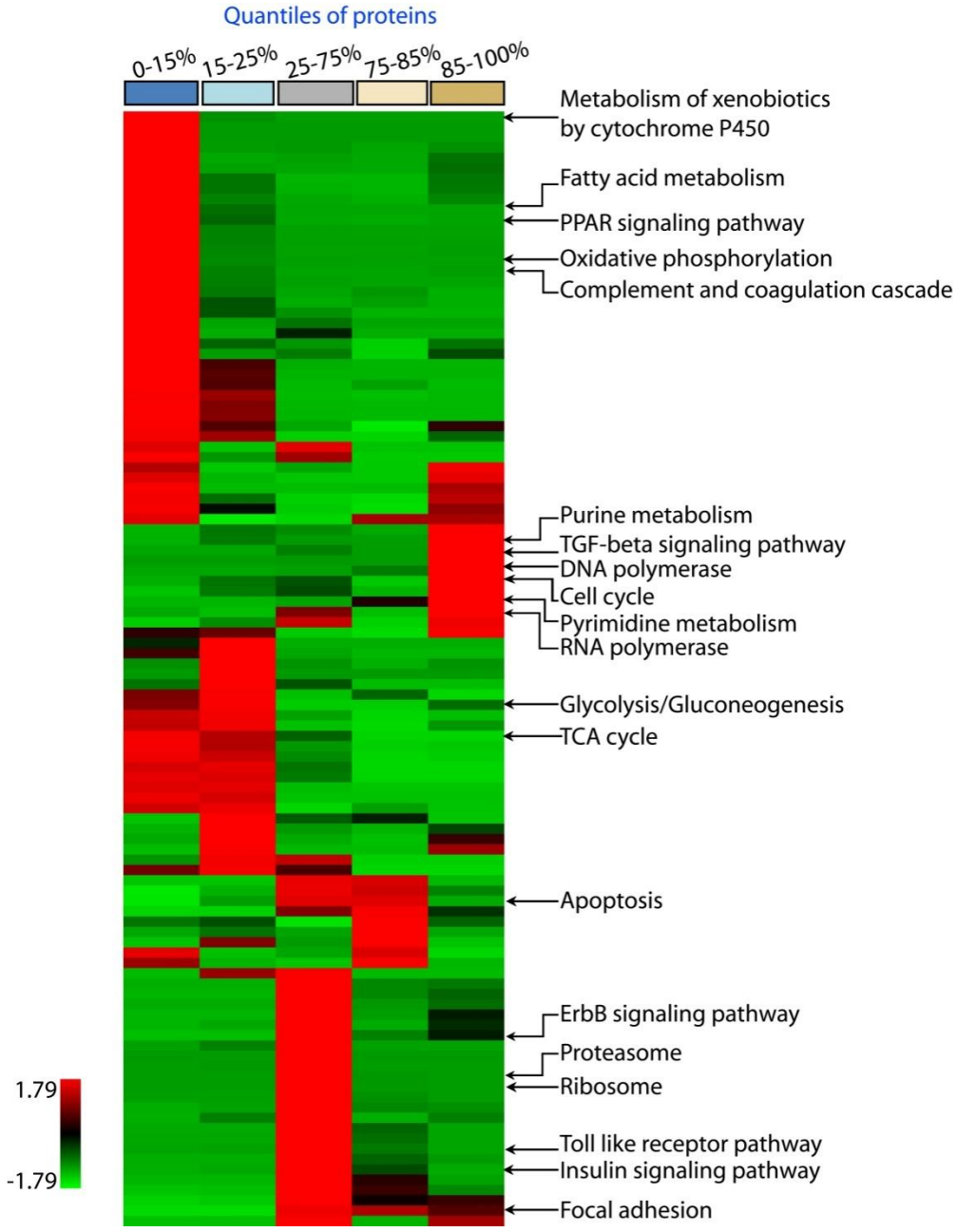


Figure 7.6 Functional phenotyping of the proteome. The five quantiles (see Figure 7.2) were separately analyzed for enriched KEGG pathways and clustered for the z-transformed p -values. The color bar on top represents the quantiles. Representative pathways enriched in the protein population of each quantile are annotated.

7.3.3 Proteomic differences between Hepa1-6 and primary hepatocytes revealed by systematic bioinformatics

The most prominent cluster of proteins expressed at higher levels in Hepa1-6 relates to cell division and encompasses categories such as cell cycle ($p < 10^{-9}$) (Figure 7.7), DNA synthesis ($p < 10^{-4}$) and RNA polymerase ($p < 10^{-3}$). This cluster consists of 10 enriched pathways, of which at least five relate to increased cell proliferation. Biologically, this is not surprising since hepatocytes in the liver and in our primary culture are largely arrested in the G_O phase of the cell cycle, whereas Hepa1-6 cells double every 18 hours. Nevertheless, the fact that this phenotypic trait is so clearly grouped in the cluster analysis makes it an excellent positive control.

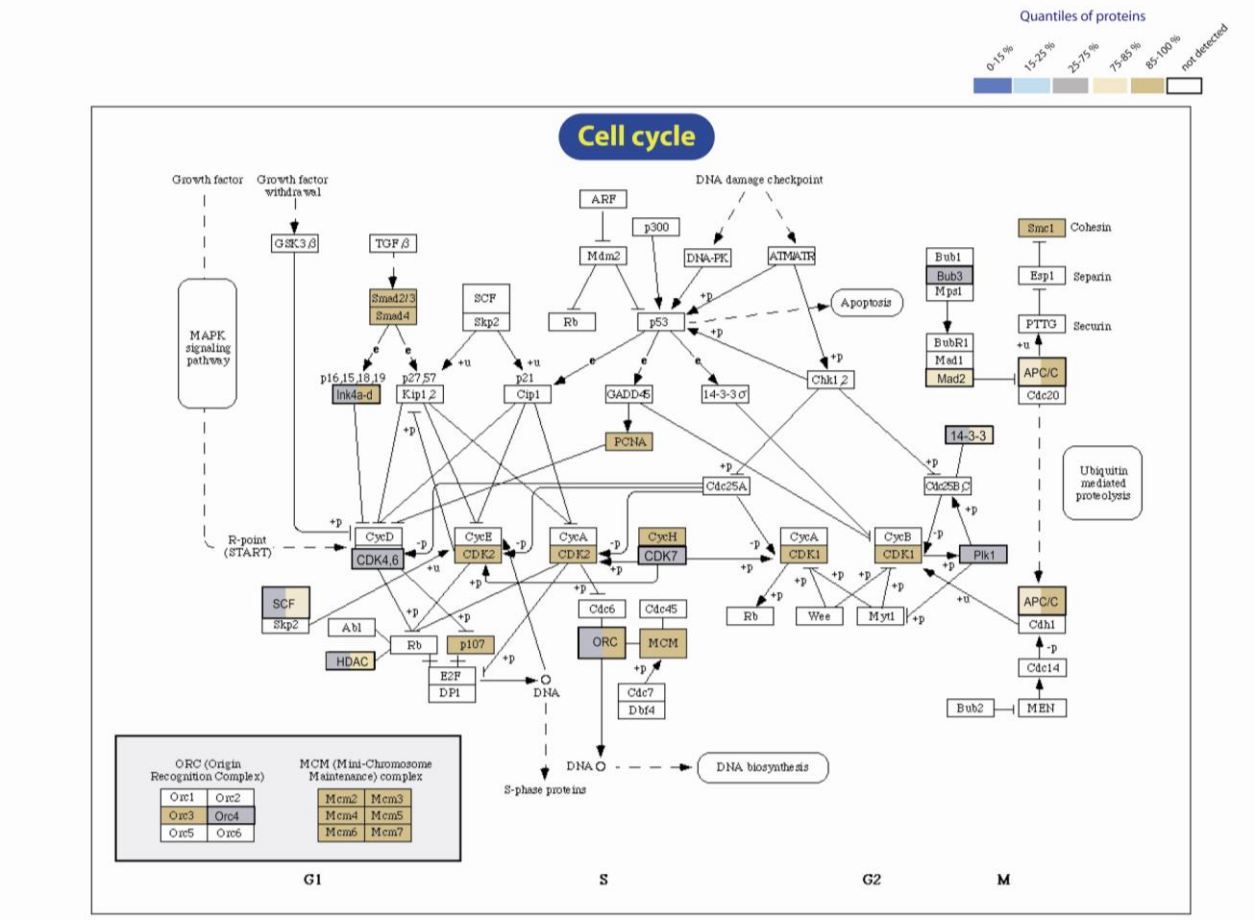


Figure 7.7 KEGG pathway mapping of cell cycle. The color bar on top represents the quantiles. Proteins that are marked with two colors are likely to represent isoforms

One of the most enriched categories in the quantile most expressed in primary cells is the P450 family of enzymes ($p < 10^{-16}$). These enzymes are mainly involved in metabolizing endogenous substances and xenobiotics[289], a prototypical function of the liver. We identified 32 different P450 proteins and 25 of them were down-regulated at least tenfold in the cell line. Furthermore, the flavin monooxygenase (FMO), UDG-glucuronosyltransferase (UGT), sulfotransferase (SULT), and glutathione S-transferase (GST) - additional prominent drug metabolizing enzyme families (DMEs) - were also severely down-regulated in Hepa1-6 (Table 7.1).

Protein Names	Uniprot ID	# of peptides	# of unique peptides	Ratio H/L Normalized	Ratio H/L Count
CYP1A2	P00186	9	5	0.07	3
CYP27	Q9DBG1	13	13	0.02	23
CYP2A12	P56593	24	21	0.04	29
CYP2A4	P15392	5	1	0.07	4
CYP2B19	O55071	6	1	0.01	3
CYP2B20	Q62397	16	11	0.09	17
CYP2C29	Q64458	11	6	0.04	12
CYP2C37	P56654	9	1	0.05	4
CYP2C40	P56657	6	6	0.03	10
CYP2C44	Q3UEM4	3	3	0.14	2
CYP2C54	Q6XVG2	9	2	0.01	1
CYP2C70	Q91W64	17	17	0.07	9
CYP2D10	P24456	12	4	0.02	16
CYP2D26	Q8CIM7	16	12	0.02	22
CYP2D9	P11714	10	6	0.02	7
CYP20	Q05421	10	10	0.03	7
CYP2F2	P33267	22	22	0.03	31
CYP2J5	O54749	3	3	0.03	3
CYP39A1	Q9JKJ9	3	3	0.07	3
CYP3A11	Q64459	4	4	0.04	3
CYP3A13	Q64464	6	6	0.06	12
CYP4A12	Q91WL5	11	11	0.07	8
CYP4F13	Q99KY6	2	2	0.18	2
CYP4F14	Q9EP75	2	2	0.02	2

Table 7.1 Protein ratios for Cytochrome P450s drug metabolizing enzymes

Only three P450s were up-regulated. Two of them (CYP1A1 and CYP2S1) are known to be regulated by the aryl hydrogen receptor[290,291]. This receptor was also more highly expressed in

Hepa1-6, providing a ready explanation for the up-regulation. The third up-regulated P450 protein (RIKEN clone E130013F06) has only been characterized on the basis of sequence homology and may have functions different from traditional P450 enzymes. Reduction of DME activity is a notorious difficulty in toxicological assays in cell lines. Toxicologists therefore attempt to stimulate liver cell lines with the aim of boosting DME activity[292]. Quantitative knowledge of the changes in the profile of DME could provide a rational basis to adapt cell systems to more closely mimic hepatocytes *in vivo*.

Another prominent and cell-specific function of hepatocytes is production of plasma proteins. Figure 3 reveals that 'complement and coagulation cascade' is specific for the primary cells ($p < 10^{-2}$). Inspection of the pathway involved shows that major liver-produced factors, such as C3, C4, MBP-C, F2, F5, A2M, Serpin A1/C1 and apolipoproteins are down-regulated more than five-fold in Hepa1-6. Thus, loss of tissue context allows the cell line to shut down this function, which is nonessential for propagation in culture. The cellular compartments most overrepresented in the primary cells are mitochondria ($p < 10^{-62}$) (Figure 7.8A) and extracellular matrix ($p < 10^{-18}$). Apparently, the cell line under-expresses proteins related to communication with stroma and with tissue maintenance. Our proteome contained a total of 479 proteins annotated as mitochondrial in GO. Of these, 69% were in the asymmetric tail of the distribution, indicating they were expressed several fold lower in Hepa1-6 cells than in primary hepatocytes. We independently confirmed this observation by DAPI and Mitotracker staining (Figure 7.8B). Indeed, primary hepatocyte nuclei were smaller whereas in these cells mitochondria were more abundant with respect to Hepa1-6. Concurrent with this, fatty acid metabolism was drastically down-regulated according to enrichment analysis of KEGG pathways (Figure 7.9A). Likewise, 'oxidative phosphorylation' ($p < 10^{-29}$; Figure 7.9B), 'urea cycle' ($p < 10^{-4}$) and 'steroid biosynthesis' ($p < 10^{-2}$; Figure7.9C) were statistically significantly enriched in the quantile most expressed in primary hepatocytes. These down-regulated metabolic functions at least partially take place in mitochondria. Conversely, parts of the glycolysis pathway were up-regulated in Hepa1-6 (Figure 7.9D). Together, our results portray a drastic metabolic rearrangement, away from oxidative metabolism in the mitochondria and towards less efficient anerobic metabolism. These findings provide evidence for the Warburg hypothesis, that cancer cells shift towards glycolytic metabolic pathways[10].

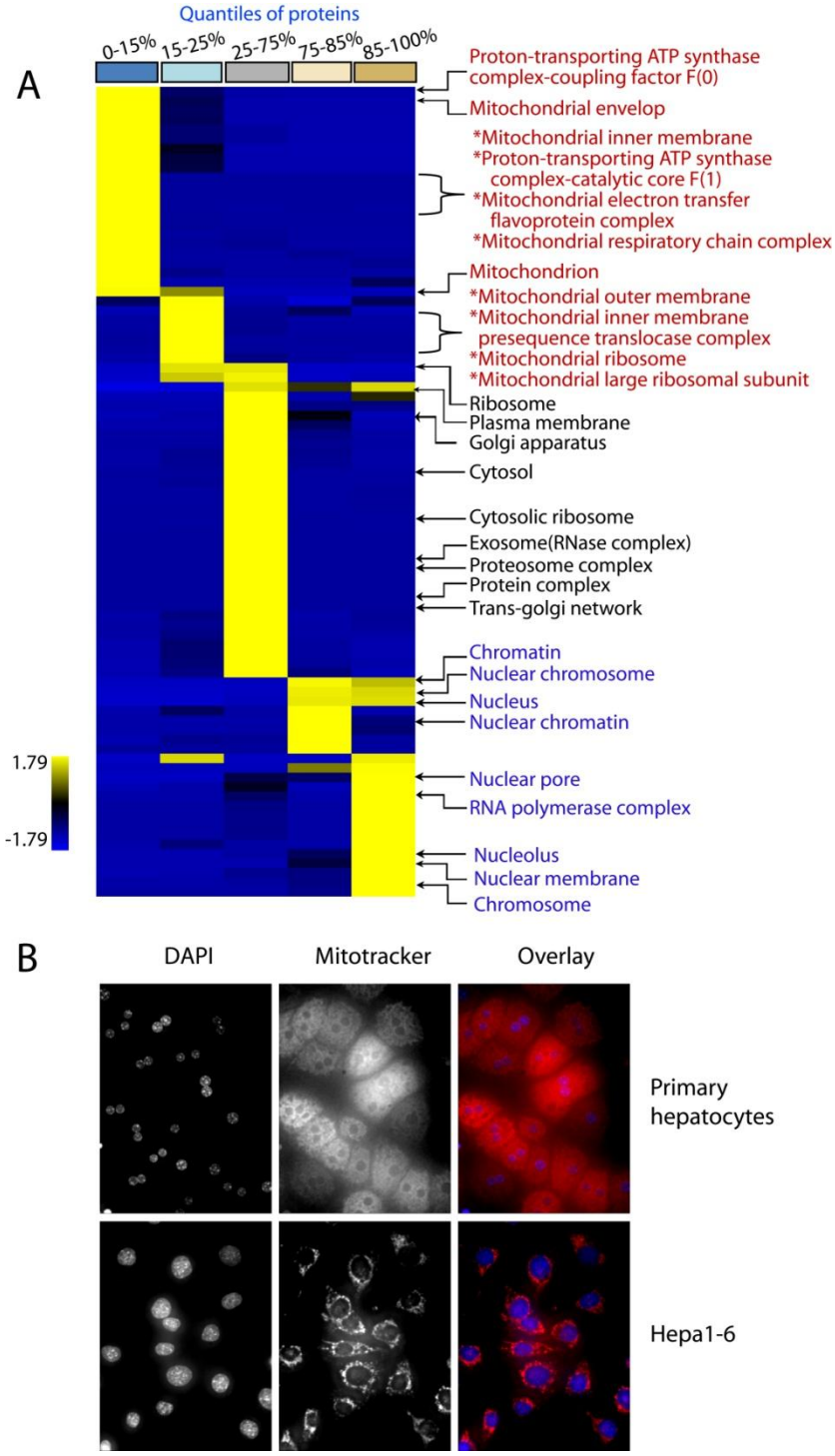


Figure 7.8 Phenotypic proteome comparison at the cellular component level. **(A)** The quantiles resulting from quantitative proteome comparison in Figure 7.3 were separately analyzed for enriched Gene Ontology Cellular Components and clustered for the z-transformed p-values. The color bar on top represents the quantiles. Representative categories enriched in the protein population of each quantile are annotated. Prominent mitochondria related categories for the primary cells are highlighted in red and prominent nucleus related categories in blue. **(B)** Nuclear (DAPI) and mitochondrial (Mitotracker) staining of primary hepatocytes and Hepa1-6 cells. Most primary hepatocytes are binuclear[293].

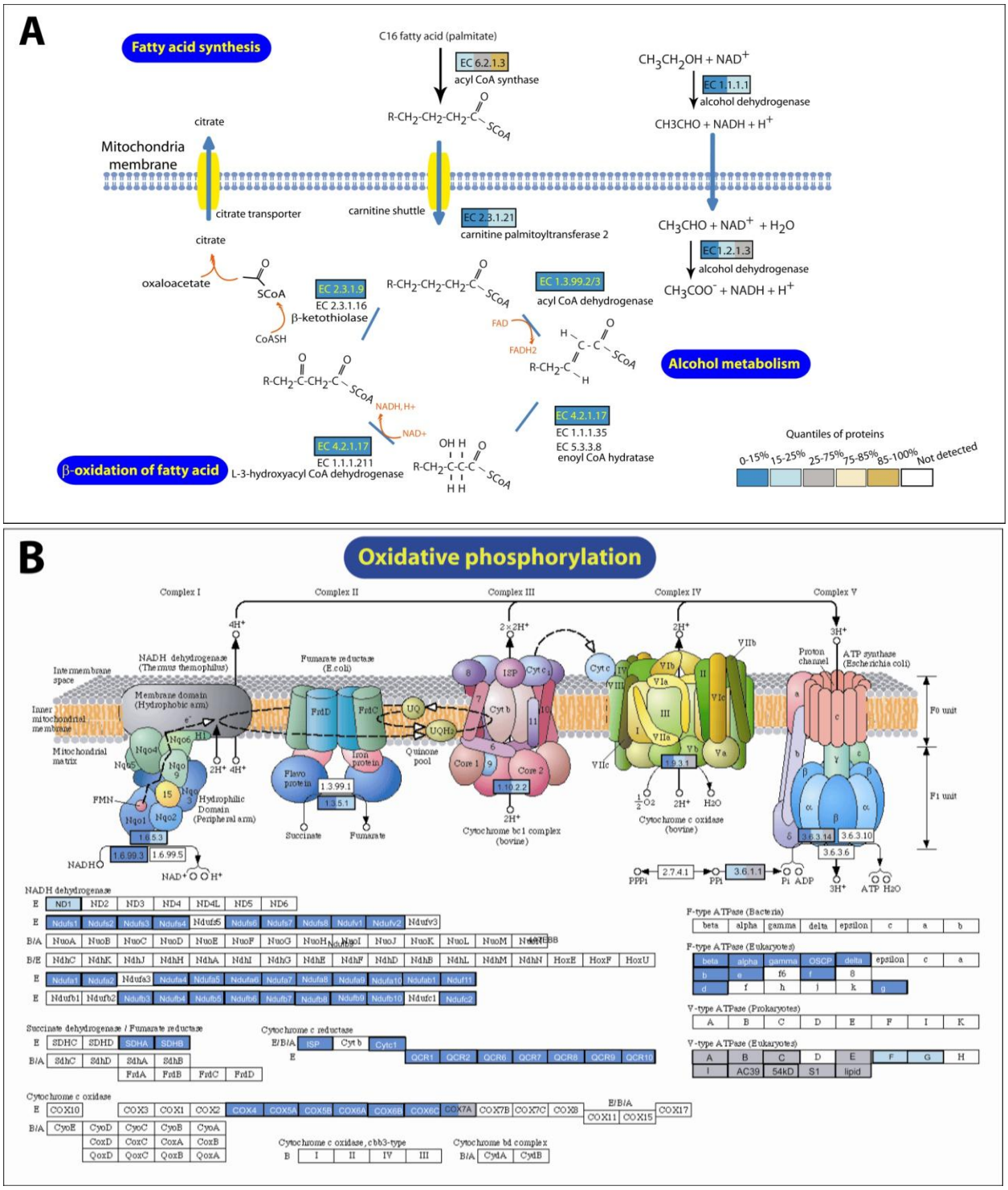


Figure 7.9 KEGG pathway mapping (A) Almost the entire fatty acid metabolism module is downregulated several-fold in Hepa1-6. Proteins are color-coded according to their relative expression in the two cell types according to the scheme in Figure 7.3. Proteins that are marked with two colors are likely to represent isoforms. **(B)** Oxidative phosphorylation, legends as in (A)

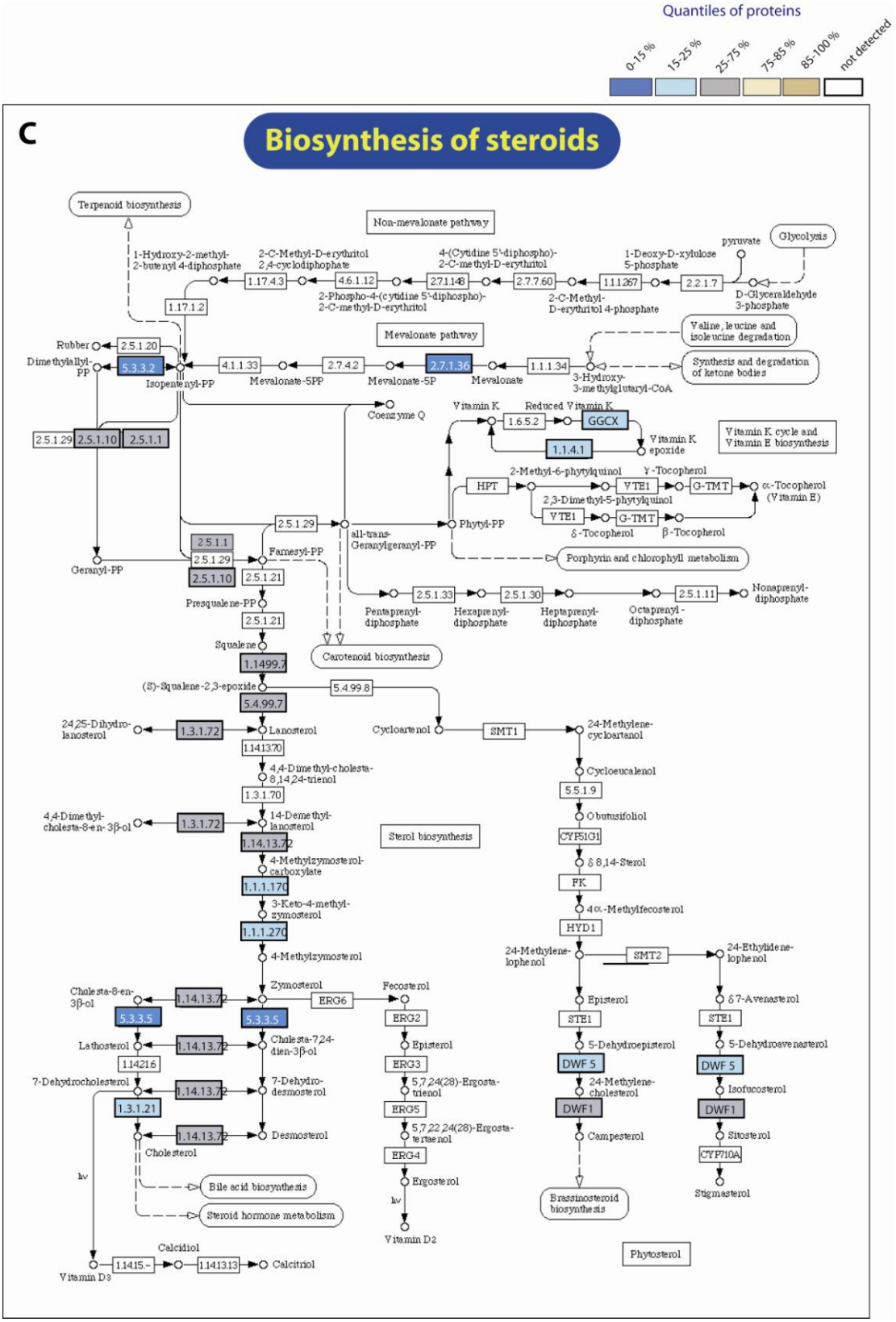


Figure 7.9 (C) Biosynthesis of steroids, legends as in Figure 7.9(A)

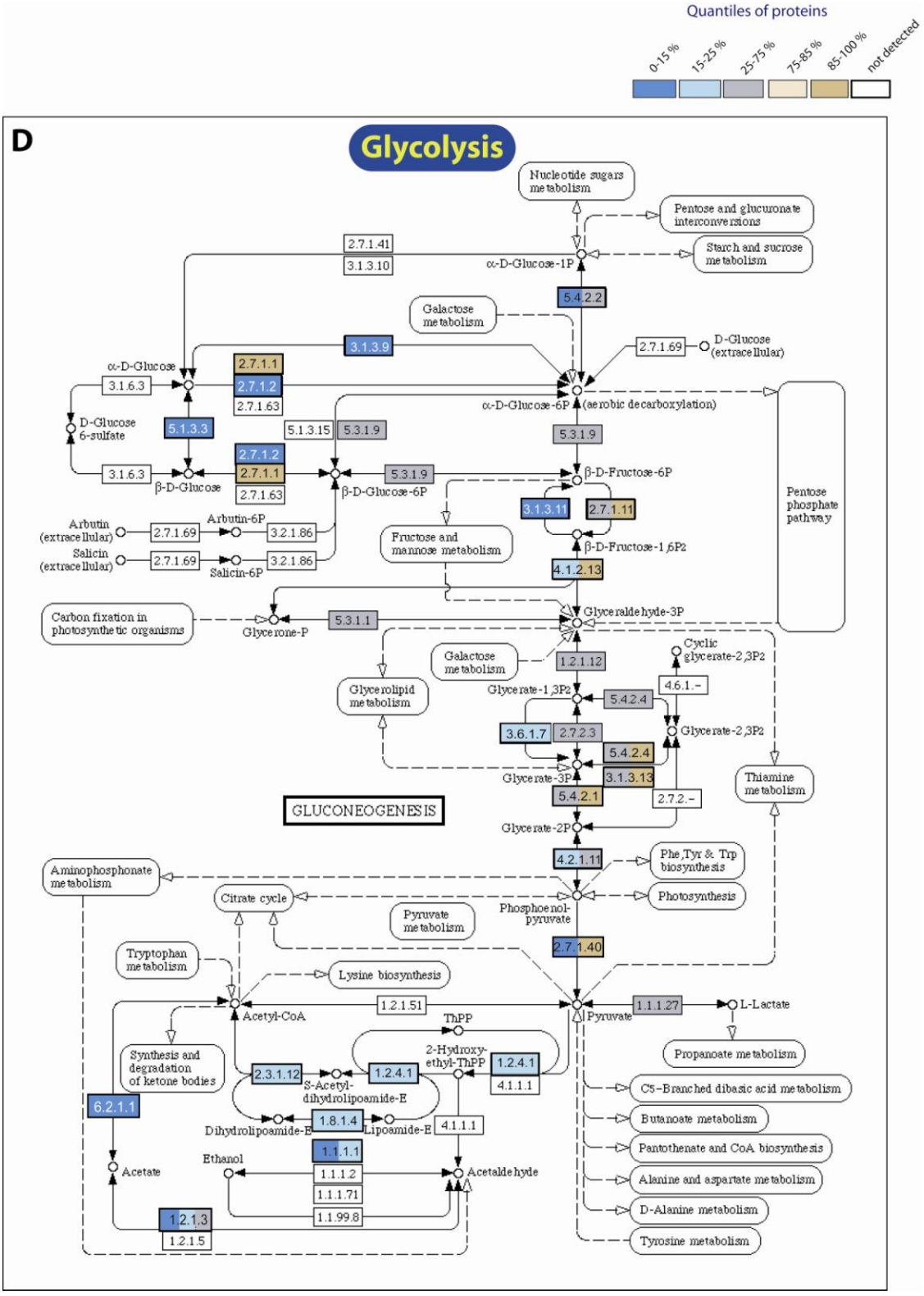


Figure 7.9 (D) Glycolysis pathway, legends as in Figure 7.9(A)

In the category containing the 50% of proteins with the least change, many household functions and organelles including ribosome ($p < 10^{-2}$), proteasome ($p < 10^{-3}$), splicing ($p < 10^{-4}$) and Golgi apparatus ($p < 10^{-3}$) are significantly enriched. Interestingly, several signaling pathways are also preferentially located in this quantile. These include the ErbB and PI3K signaling pathways (Figure 7.10).

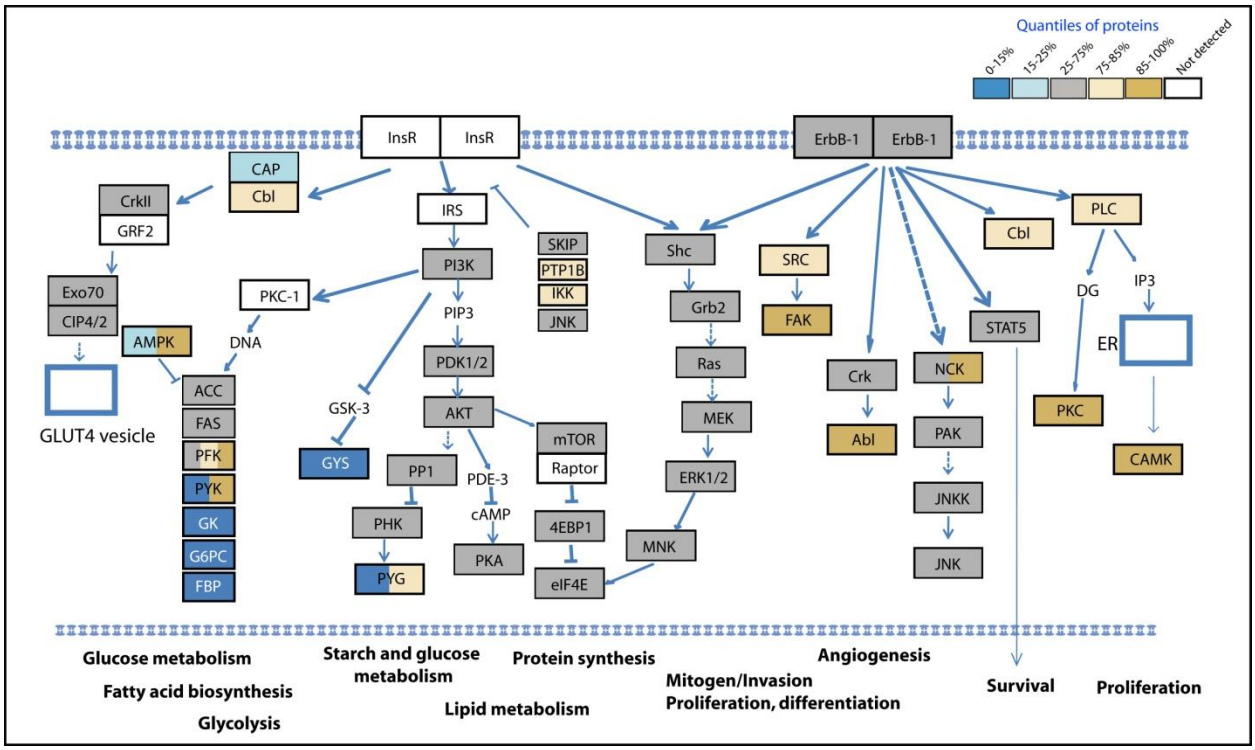


Figure 7.10 Phenotypic proteome comparisons at the pathway level. KEGG pathway mapping of ErbB and PI3K signaling pathway shows that they are equally present in primary cells and the cell line.

This finding is in agreement with the requirement of growth factor containing serum for the maintenance of most cell lines. Conversely, TGFβ-mediated signaling is more highly represented in the Hepa1-6 cell line and the canonical members TGFβ R1, Smad2/3, Smad4, p107 and p15 are all up-regulated significantly (Figure 7.11). This was unexpected because TGFβ is usually associated with growth inhibition whereas Hepa1-6 has an increased proliferation rate compared to primary hepatocytes. However, the biological actions of TGFβ are complex and it is thought to shift from a growth inhibitory to a growth promoting role during cancer development[294]. Thus up-regulation of this pathway suggests that in the Hepa1-6 tumor cells, TGFβ may have growth

promoting effects. Taken together, our data indicate that biological functions related to many important signaling pathways are well preserved in Hepa1-6.

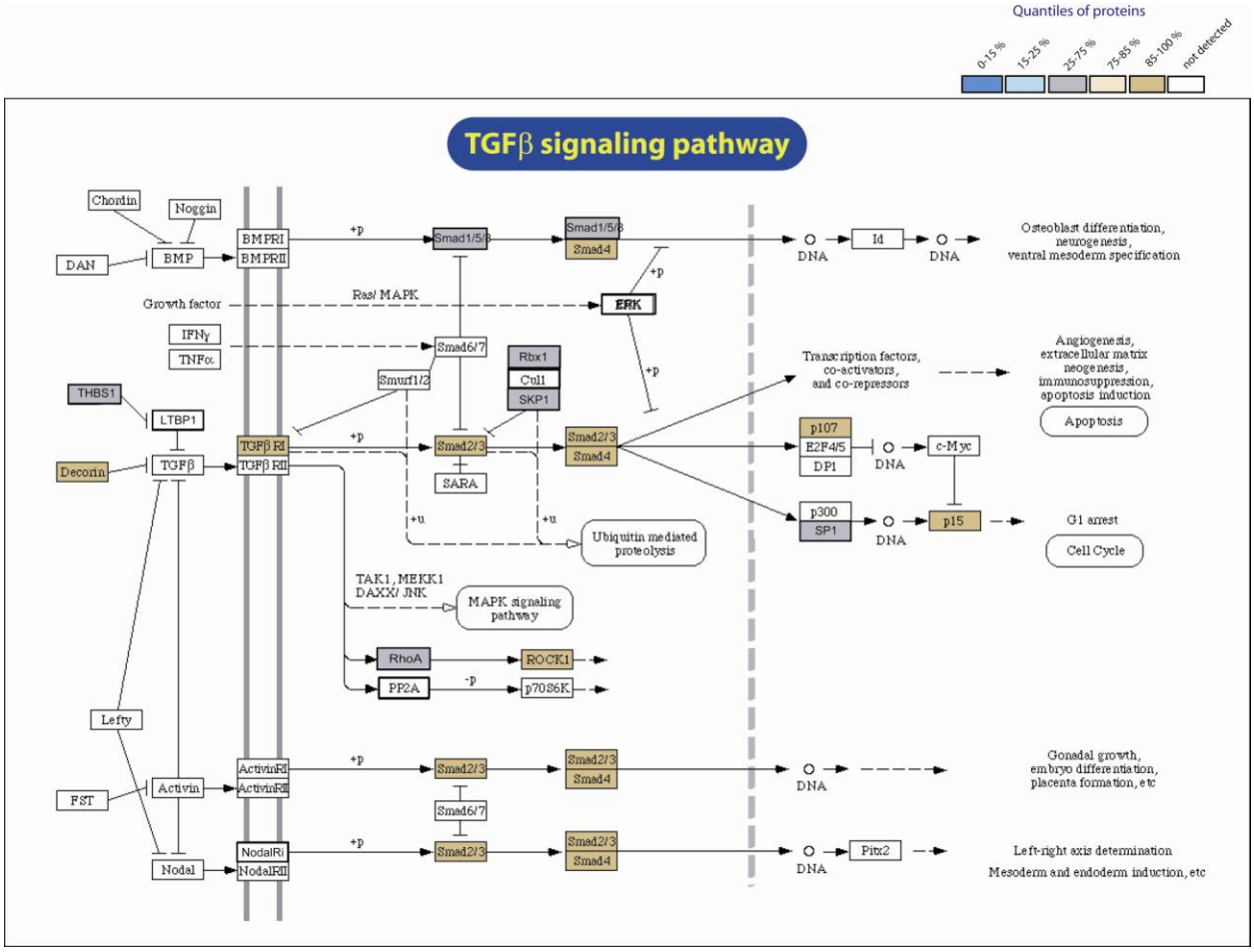


Figure 7.11 Proteomics phenotyping at the pathway level. KEGG pathway mapping shows that TGFβ signaling pathway is predominantly present in the cell line.

Some categories shared by both cell types and enriched when analyzed using the KEGG database represent non-liver functions (such as 'long term potentiation') or even non-animal functions (such as 'CO_2 fixation'). However, the enzymes found in these categories function both in liver tissue as well as in neurons or plants. Therefore, overrepresentation of these categories reflects the still evolving state of annotation of pathway databases rather than a limitation of our technology.

7.4 Discussion

Taking advantage of the ability of SILAC to compare the levels of thousands of proteins in different cellular states[2,283] and a novel bioinformatics approach, we have, for the first time, compared the proteomes of primary cells to cell lines. The overall picture that emerges is that Hepa1-6 has lost many of the specific functions typical of hepatocytes *in vivo*. Examples are the DMEs, complement production and synthesis of extracellular matrix. Conversely, the cell line shifts more of its resources into functions associated with proliferation, but maintains important cell signaling pathways. This phenotype is 'rational' for rapidly dividing and not nutrient limited cells and may partly reflect Darwinian selection of cell clones.

Our technology is accurate, relatively rapid and should now allow selection of the appropriate cell system based on a global and unbiased profile according to desired biological function. Furthermore, it can be used to manipulate the cell line system to better reflect the *in vivo* situation at the proteome level. While we have based our analysis on protein expression levels, it could just as well be applied to assess fidelity of signaling pathways in cell lines using SILAC-based quantitative and global phosphoproteomics[65].

Our bioinformatics analysis differs in important points from the more familiar measurement of mRNA levels by microarray and its associated bioinformatics[295]. Even though reproducibility of microarray chips has become much better during recent years, the data is not quantitative with respect to the final, desired parameter – the global change in protein levels. Furthermore, results of any specific transcript on the chip generally have to be validated by RT-PCR and then by quantitative immunoblotting. This is impractical for large numbers of proteins. In contrast, quantitative proteomics inherently contains the fold-change for each protein, and increasingly also that of specific isoforms. The quantitative nature of our results also made it possible to directly group overrepresented functions and processes instead of the genes themselves.

Here we have analyzed interesting, but relatively general phenotypic traits of two cell populations. While many of the resulting observations can be immediately rationalized in terms of biological function, they have never been quantified in a global and unbiased way. Our data

furthermore contains a wealth of functional leads that could not be explored in depth here. The combination of very high quantitative accuracy at the proteome level with increasingly accurate pathway databases should allow even richer assessment of the phenotypic state of any cell population in the future.

8. A systems view of the cell cycle by quantitative phospho-proteomics

This work is included in a manuscript under submission:

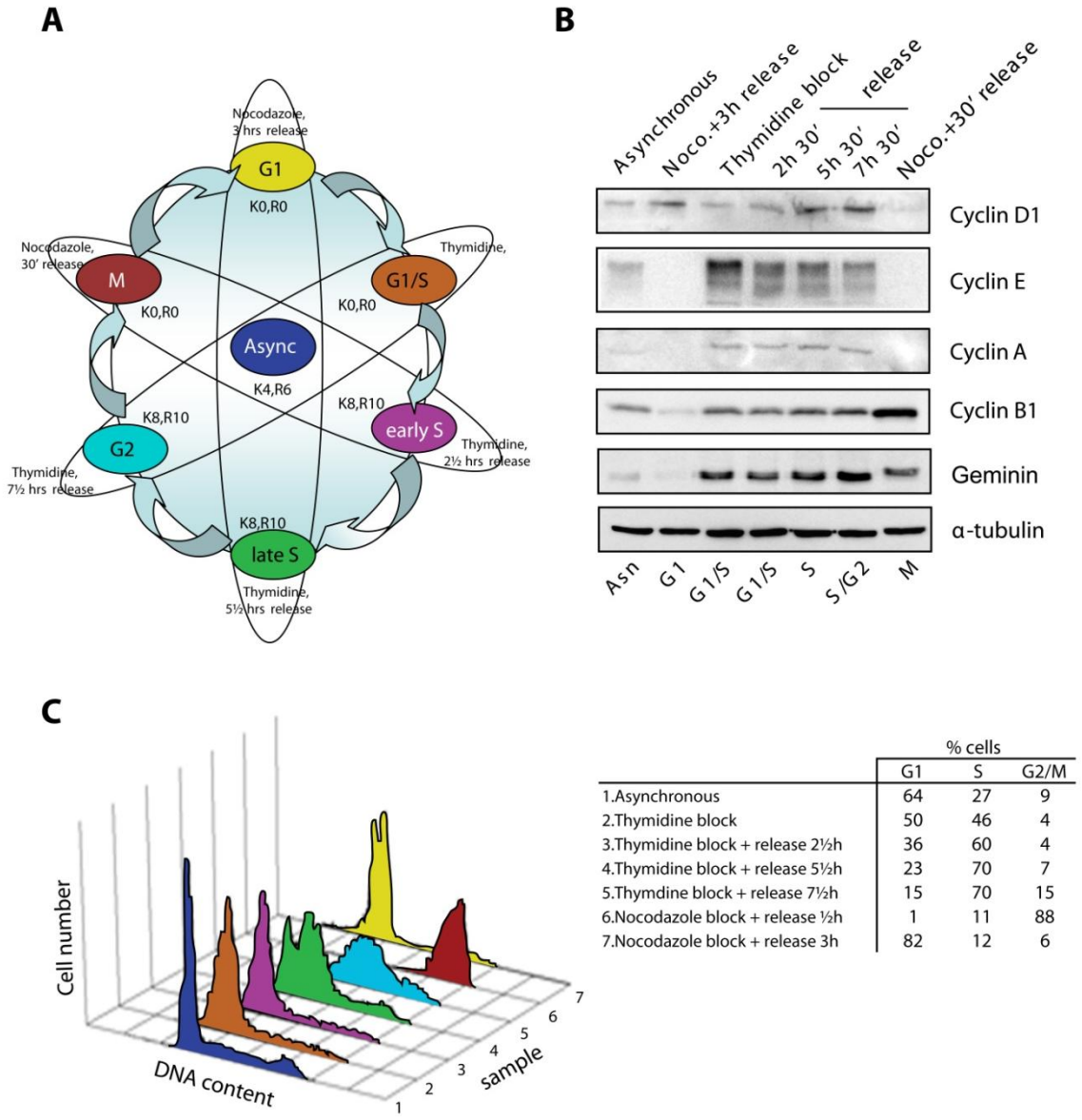
Jesper V. Olsen[φ], Michiel Vermeulen[φ], Anna Santamaria[φ], **Chanchal Kumar**[φ], Martin L. Miller, Lars J. Jensen, Florian Gnad, Juergen Cox, Thomas S. Jensen, Erich A. Nigg, Søren Brunak, Matthias Mann

A systems view of the cell cycle by quantitative phosphoproteomics

[φ] These authors contributed equally to this work

8.1 Introduction

The cell cycle is a highly regulated and evolutionarily conserved process that results in the duplication of the cell's content and involves a large number of dedicated protein complexes and signaling pathways. Deregulation of key players that coordinate this process is implicated in tumorigenesis[296,297]. Global analyses of the cell cycle have so far been limited to microarray studies of the transcriptome[298]. However, regulated protein phosphorylation and protein degradation both play pivotal roles in controlling the cell cycle, but these are not always directly reflected by mRNA changes. Recent advances in proteomics technology, in particular high-resolution mass spectrometry (MS), now allow large-scale quantitation of the proteome and the phosphoproteome. For example, we have recently quantified more than six thousand phosphorylation sites in response to cell stimulation[65] using a metabolic labeling technique termed SILAC[9]. Here we combine this technology with global proteome quantitation to obtain a systems-biology view of protein and phosphorylation dynamics during the human cell cycle. We SILAC-encoded three HeLaS3 cell populations using light, medium and heavy stable isotopic versions of arginine and lysine. Light and heavy populations were synchronized in six different stages of the cell cycle and mixed with medium labeled asynchronously growing cells (Figure 8.1A). Because of their different molecular weights, peptides from the three populations are separated in the mass spectrometer and are directly quantifiable against each other.



	% cells		
	G1	S	G2/M
1.Asynchronous	64	27	9
2.Thymidine block	50	46	4
3.Thymidine block + release 2½h	36	60	4
4.Thymidine block + release 5½h	23	70	7
5.Thymdine block + release 7½h	15	70	15
6.Nocodazole block + release ½h	1	11	88
7.Nocodazole block + release 3h	82	12	6

Figure 8.1 Quantitative proteomic analysis of the human cell cycle. **(A)** HeLa S3 cells were SILAC labeled with three different isotopic forms of arginine and lysine. Three individual populations of heavy and light SILAC cells were pre-synchronized using a thymidine-block and then collected at six different time-points across the cell cycle following release from the thymidine arrest. Two samples were collected after an additional nocodazole-arrest and release. Cells were lysed and mixed in equal amounts using an asynchronously growing cell population (medium SILAC) as the internal standard allowing normalization between experiments. Three independent experiments were performed to cover six cell cycle stages. **(B)** Immunoblot analysis of known cell cycle marker proteins in the different cell populations. **(C)** Fluorescence Activated Cell Sorting (FACS) profiles of the individual synchronized HeLa S3 populations. Cells were fixed, collected by centrifugation after which the DNA content of the cells was determined using propidium iodide.

This results in relative quantitation of the proteome and the phosphoproteome during the cell cycle in three experiments with the asynchronous population present in each experiment for normalization. Following harvesting, cells from the three experiments were combined and processed as described in section 8.2 for proteome and phosphoproteome analyses. We monitored synchronization of SILAC cell cultures by Western blotting and FACS analyses (Figure 8.1B and 8.1C).

8.2 Materials and Methods

8.2.1 Cell culture and sample preparation

HeLa S3 cells were SILAC labeled as previously described using three different isotopic versions of lysine and arginine[9]. Cells were synchronized in G1/S overnight using a thymidine block at a concentration of 4 mM (Sigma, St. Louise, MO). Cells were then released from thymidine block and subsequently collected at four different time points; 0 h (G1/S-phase), 2.5 h (Early S-phase), 5.5 h (Late S-phase) and 7.5 h (G2-phase) after removal of thymidine. Two sets of cells were arrested overnight using nocodazole following the 7.5 hours release from thymidine. The next morning these cells were released for either 0.5 h (M-phase) or 3 h (G1-phase). Western blotting and FACS analyses were performed to monitor the efficiency of the cell-cycle arrest (Fig. 8.1B,C).

8.2.2 Fluorescence-activated Cell Sorting Analysis

Cell suspensions were fixed with 80% ethanol, permeabilized by treatment for 5 min with 0.25% Triton X-100 in PBS, and incubated with 0.1% RNase and 10 µg/ml Propidium Iodide. Cellular DNA content was determined by flow cytometry using FACSCalibur (BD Biosciences Clontech, San Jose, CA) system and CellQuest software (Becton-Dickinson, Lincoln Park, NJ).

8.2.3 Western blotting

Cells were washed once with ice-cold PBS containing 1 mM phenylmethylsulfonyl fluoride, scraped off the plate, and resuspended in ice-cold HEPES lysis buffer (50 mM HEPES, pH 7.4, 150 mM NaCl, and 0.5% Triton X-100) containing 1 mM DTT, 30 µg/ml RNase A, 30 µg/ml DNase, protease, and phosphatase inhibitors. After 15 min on ice, lysed cells were centrifuged at 13,000 rpm for 15 min at 4°C. Protein concentrations in the cleared lysate were determined using

the Dc protein assay (Bio-Rad), and equal protein amounts were loaded on SDS-PAGE gels. Separated proteins were transferred to nitrocellulose membranes (Whatman Schleicher and Schuell). For Western blot analysis, rabbit anti-cyclin D (St. Cruz Biotechnology), mouse mAb anti-cyclin E (clone HE-12 tissue culture supernatant), goat anti-cyclin A (St. Cruz Biotechnology), mouse anti cyclin-B (BD Transduction Laboratories), mouse mAb anti-α-tubulin (Sigma-Aldrich), rabbit anti-Geminin (gift from Roland A. Lasckey), rabbit mouse anti-Bub1 (clone 61–22–2, tissue culture supernatant), rabbit anti-Eg5[299], mouse anti-Aurora B (BD Transduction Laboratories), mouse anti-Plk1 (clone PL2, tissue culture supernatant), mouse anti-securin (Abcam), mouse anti-TPX2 (Abcam), rabbit anti-Kif20A (gift from Thomas U. Mayer), mouse anti-pT210 Plk1 (BD Transduction Laboratories), anti-pT14 Cdk1, rabbit and rabbit anti-pS10 Histone 3 (Upstate Biotechnology) were used and detected by ECL Supersignal (Pierce Chemical) using a digital Fujifilm LAS-1000 camera attached to an Intelligent darkbox II (Raytest).

Arrested cells were lysed in modified RIPA buffer[65] after which protein extracts were clarified by centrifugation to pellet chromatin and other insoluble material. This insoluble pellet was redissolved in 8M urea/1% N-octylglucoside supplemented with phosphatase inhibitors and benzonase. The soluble proteins in the RIPA extract were precipitated overnight at -20 °C by adding 4 volumes of ice-cold acetone. Following centrifugation, precipitated proteins were redissolved in 8M urea/1% N-octylglucoside supplemented with phosphate-inhibitors. The protein concentration of all the fractions was determined using the Bradford assay. Protein extracts derived from the different cell-cycle arrest stages were then mixed 1:1:1 accordingly using an asynchronous cell population as the internal standard. 20% of the protein mixtures were separated by 1D-SDS PAGE, sliced in 20 gel-plugs and digested with trypsin in-gel[300]. 30% of the extracted peptide mixtures were used for quantitative proteome analysis by LC-MS, whereas the other 70% of the extracted peptides were subjected to titanium dioxide enrichment in the presence of 2,5-DHB[38] and analyzed by LC-MS. The remaining 80% of protein mixtures were not fractionated by 1D SDS PAGE but directly reduced with DTT, alkylated using iodoacetamide and subsequently digested with endoproteinase Lys-C and trypsin as described[65]. The resulting peptide mixtures were either directly subjected to titanium oxide enrichment or first fractionated by strong-cation chromatography followed by titanium dioxide enrichment.

8.2.4 Mass Spectrometry

All experiments were performed on an LTQ-Orbitrap instrument connected to an online nanoflow HPLC (Agilent 1100 system) via a nanoelectrospray ion-source (Proxeon Biosystems). The tryptic peptide mixtures were autosampled onto a 15 cm long 75 µm ID column packed in-house with 3-um C18-AQUA –Pur Reprosil reversed-phase beads (Dr. Maisch) and eluted with a linear gradient from 8% to 40% MeCN in 2 hrs. The separated peptides were electrosprayed directly into an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific), which was operated in the data-dependent acquisition mode to automatically switch between one orbitrap full-scan and five ion trap tandem mass spectra. The tandem mass spectra were acquired with the multi-stage activation enabled for neutral loss of phosphoric acid (32.66, 48.99 and 97.97 amu)[301]. All full-scan spectra were recalibrated in real-time using the lock-mass option[86].

8.2.5 Data processing and analysis

Mass spectrometric data were analyzed using the in-house developed software MaxQuant version 1.0.12.0[2]. Which performs peak list generation, SILAC- and XIC-based quantitation, estimation of false discovery rates for search engine results, peptide to protein group assembly, as well as data filtering and presentation. The MS/MS spectra were searched against the human International Protein Index sequence database (IPI version 3.37) supplemented by frequently observed contaminants, concatenated with reversed versions of all sequences. Mascot (version 2.2.04) was used for the database search. Enzyme specificity was set to trypsin, allowing for cleavage N-terminal to proline and between aspartic acid and proline. Carbamidomethyl cysteine was set as fixed and oxidized methionine, N-acetylation, loss of ammonia from N-terminal glutamine as well as phosphorylation of seine threonine and tyrosine as variable modifications. Spectra determined to result from medium or heavy labeled peptides by pre-search MaxQuant analysis were searched with the additional fixed modifications Arg6 and Lys4 or Arg10 and Lys8, respectively, while spectra with a SILAC state not determinable a priori were searched with Arg10 and Lys8 as additional variable modifications. A maximum of three missed cleavages and three labeled amino acids (arginine and lysine) were allowed. The required false discovery rate was set to 0.01 at the peptide and at the protein level and the minimum required peptide length to 6 amino acids. If the identified peptide sequence set of one protein was equal to or contained another protein's peptide set, these two proteins were grouped together by

MaxQuant and not counted as independent protein hits. Protein SILAC ratios are reported as the median of the ratios derived from SILAC triplets assigned to the protein. For phosphopeptides the phosphorylation site(s) were assigned by using a modified version of the PTM score[65] in MaxQuant. All high-confidence phosphosites (FDR<0.01) together with their cell cycle dependent ratios were uploaded to Phosida (http://www.phosida.com), which is a freely accessible phosphorylation site repository[302].

8.2.6 Peak time index calculation for (phospho)-proteomic temporal profiles

The workflow of the analysis is shown in Figure 8.3. The fold ratios (r_1 through r_6) for each protein over the 6 time points (t_1=1 through t_6=6) were scaled between range [0, 1]. Then for each protein we calculated a time peak index (t_{peak}) by weighted mean of the expression ratio of maximal expression (i.e. r_i=1) at time point t_i with respect to its adjacent time points (t_{i-1} and t_{i+1}). In order to maintain the cyclicity we made two assumptions: (a) if the maximal expression was at t_1 (i.e. r_1 = 1) then t_1 was preceded by t_0=0 with expression r_6, and (b) if the maximal expression was at t_6 (i.e. r_6 = 1) then t_6 was followed by t_7=7 with expression r_1. The equations for the peak time (t_{peak}) calculation are as follows:

$$t_{peak} = \begin{cases} \dfrac{t_{i-1} * r_{i-1} + t_i * r_i + t_{i+1} * r_{i+1}}{r_{i-1} + r_i + r_{i+1}}, if \max(r_i)\ at\ i \in [2,5] \\ \dfrac{t_i * r_i + t_{i+1} * r_{i+1} + 0 * r_6}{r_i + r_{i+1} + r_6}, if \max(r_i)\ at\ i = 1 \\ \dfrac{t_{i-1} * r_{i-1} + t_i * r_i + 7 * r_1}{r_{i-1} + r_i + r_1}, if \max(r_i)\ at\ i = 6 \end{cases}$$

The protein expression profiles were subsequently ordered in increasing order to get a temporal map of cell cycle. For the purpose of rendering according to their increasing t_{peak} the original (unscaled) expression profile for each protein was *z*-transformed prior to rendering as in Figure 8.4.

8.2.7 Cyclic angular peak calculations based on peak time index of (phospho)-proteomic temporal profiles

The time peak measure for any protein j, $t_{peak(j)}$ was further converted to an angular peak measure $\theta_{peak(j)}$ in the range [0,360°] by following equation:

$$\theta_{peak(j)} = \frac{t_{peak(j)} - \min_{k\in[1,N]}(t_{peak(k)})}{\max_{k\in[1,N]}(t_{peak(k)}) - \min_{k\in[1,N]}(t_{peak(k)})} * 360° \ , N = \ number\ of proteins\ in\ dataset$$

Thus, the time peak measures $t_{peak(j)}$ were converted to a polar coordinate system with the radial coordinate r=1 and the polar angle $\theta_{peak(j)}$. By definition the polar coordinate system has an anticlockwise orientation with 0° ray as the polar axis. However, in order to represent and analyze our proteomics data according to the standard cell cycle stages beginning with "Mitosis" we further wished to choose a transformed polar coordinate system having clockwise orientation with 90° ray as the polar axis. In such a polar coordinate space $\theta_{peak(j)}$ has to be transformed by following equation: $\theta^*_{peak(j)} = 90^o - \theta_{peak(j)}$ .These $\theta^*_{peak(j)}$ values were used to render the *z*-transformed protein profiles in the transformed polar coordinate space as shown in Figure 8.8.

8.2.8 Enrichment analysis for Gene Ontology Cellular Component (CC) based on circular statistics

For each of the Gene Ontology[288] cellular component category ***C*** the circular peak angles of the complete protein set (*N* proteins) was used to derive a 1×N vector θ_C such that its j^{th} entry was $\theta^*_{peak(j)}$ if protein *j* was annotated with ***C*** else "NA". This θ_C vector was tested for non-homogeneous distribution across the unit circle in the transformed polar coordinate system (θ^*) by using the "Rayleigh test"[303]. Only categories which had a *p*-value < 0.05 were considered significant. The mean(θ_C) provided mean direction of enrichment for the category ***C*** and was used to render the category at particular angles in Figure 8.8.

8.2.9 Comparison with cell cycle microarray dataset

The microarray dataset of ref[298] experiment no. 4 (Thy-Noc) was chosen for comparison with our proteomics dataset as the experimental conditions therein paralleled our study. The complete microarray dataset was categorized into changing (1,100 probes) and non-changing (39,484 probes). The complete identified proteome was divided into regulated (2,857 IPIs) and non-regulated (3,169 IPIs) proteins. The IPI identifiers of the proteomics data were mapped to the microarray probes using common EntrezGene identifiers. Some of the IPIs could be mapped to more than one EntrezGene identifiers and hence were multiplicatively mapped for each identifier

thereby resulting in 11,826 probe entries in common. This final mapped dataset was categorized into 4 classes as shown in following contingency table:

		Proteins	
		Regulated	**Non-regulated**
mRNA	**Cycling**	Cluster A(311)	Cluster B(226)
	Non-cycling	Cluster C(5,493)	Cluster D(5,796)

These set of clusters were then analyzed using GO based clustering method for enriched biological processes (BP) by a method similar to the one described in section 8.2.11.

8.2.10 Comparison with steady-state HeLa microarray data

We used a recently published microarray data set[304], which employed the Affymetrix HGU133A GeneChip with 22,283 probe sets that map to 12,999 Entrez gene identifiers. We downloaded the data from the four control data sets representing the normal HeLa transcriptome. In accordance with practice in our proteomics experiment, we defined a transcript as present if the MAS5 p-values were at least 0.01 in three out of the four experiments (the MAS5 probability values are a standard measure of the presence of a transcript, and they are calculated from the signals of the different elements in each probe set). We then mapped the 8,161 probe sets with a "present call" on the Affymetrix probe set to 5,791 unique Entrez gene identifiers. Our quantified proteome of 6,026 IPI entries was mapped to 5,455 unique Entrez gene identifiers. Subsequently the overlap between the two datasets was calculated by common Entrez gene identifiers. Figure 8.2 shows their overlap with the proteomic HeLa data set.

8.2.11 Gene Ontology and KEGG pathways enrichment based clustering for protein groups based on peak time

The enrichment analysis for Gene Ontology (GO)[288] Biological Process(BP) and Cellular Component(CC) were done separately for each of the peak clusters (M peak, G1 peak, G1/S peak, Early S peak, Late S peak, G2 peak) derived from peak time index clustering (Figure 8.4) with respect to the whole quantified proteome by the "conditional hypergeometric test" available in the GOstats[284] package in the R statistical environment[305]. For further hierarchical clustering based on GO terms we first collated all the categories obtained after enrichment along with their

p-values, and then filtered for those categories that were at least enriched in one of the clusters with p-value < 0.05. This filtered p-value matrix was transformed by the function $x = -\log_{10}$ (p-value). Finally these x values were z-transformed for each GO category. These z-scores were then clustered by one-way hierarchical clustering (Euclidean distance, Average Linkage clustering) using Genesis[241]. KEGG pathway[306] enrichment analysis was done in the same way, except that the 'hypergeometric test' was employed and the reference set were the complete human KEGG annotations.

8.2.12 Analysis of kinase–substrate relationships during phases of the cell cycle

In order to predict kinase-substrate relationships all identified class I serine and threonine phosphorylation sites (pS/T) were scored with NetPhosK[307]. We found that the overall score distribution of the cycling pS/T is significantly different from the score distribution of the non-cycling counterparts ($P < 10^{-10}$, Chi-square test). For each time point in the cell cycle we investigated which kinases contribute most to the observed differences in score distributions compared to non-cycling pS/T. This was plotted with the "heatmap" package in R (the time points in the cell cycle were z-score scaled) (Figure 8.11C).

8.2.13 New candidates in the DRR network

We found 479 pS/T sites in the data set that matched the pS/T-Q DNA Damage Repair (DDR) kinase consensus motif. Out of these sites we defined a dynamic subset that was regulated on multiple levels. This subset consisted of 13 sites regulated on phosphorylation (high confidence) and protein levels as well as 8 sites regulated on phosphorylation (intermediate confidence) and mRNA levels. We next used the NetworKIN algorithm[308] to classify which upstream kinases target the particular phosphorylation sites and found that 15 pS/T-Q sites in 14 proteins were contextually linked to ATM or DNA-PK.

8.3 Results

8.3.1 High throughput identification of proteome changes during the cell cycle

LC-MS/MS data from 144 gel slices for the quantitative proteome and 300 fractions for the phosphoproteome were together analyzed using the MaxQuant software[2] applying unified statistical criteria. At a false discovery rate of one percent this resulted in the identification of 6,695 proteins, of which 6,281 were directly associated to 5,878 unique Ensembl genes. For 6,027 proteins, quantitative profiles were obtained. About 70% of the proteome was phosphorylated (4,795) and a total of 23,765 phosphorylation sites were identified – again at a false discovery rate of one percent. We could confidently assign 18,037 unique phosphorylation sites in the peptide sequences (Class I sites[65]). As our phosphoproteome measurement is extensive but still not complete, our data suggests that the majority of all human proteins are phosphorylated to some degree.

8.3.2 Coverage of the proteome

To assess if our measured HeLa cell proteome is biased against 'difficult' protein classes such as low-abundance regulatory proteins or membrane proteins and determined to which extent we covered these categories and well-known protein complexes. Typically we identified at least 70%, suggesting the HeLa cell proteome contains at least 10,000 proteins of which we quantified a majority. Classical cell-cycle proteins, such as cyclins, CDKs and components of the anaphase promoting complex/cyclosome (APC/C) were measured essentially completely. This is the largest quantified proteome to date, similar in size to typical dynamic transcriptome profiles of human cell lines (see section 8.2.10, Figure 8.2). The quality of the proteome-wide quantitation is demonstrated by the dynamic MS-based profiles for the marker proteins shown by Western blotting in Figure 8.1B and other key cell-cycle proteins. Expression levels of a fifth of the proteome changed by at least four-fold over the cell cycle. A four-fold change also best accounted for the dynamics of already described cell-cycle components. A recently published global study using RNAi identified a cell-cycle phenotype for more than 1,000 proteins[309]. Both numbers are large compared to the number of known central cell-cycle actors.

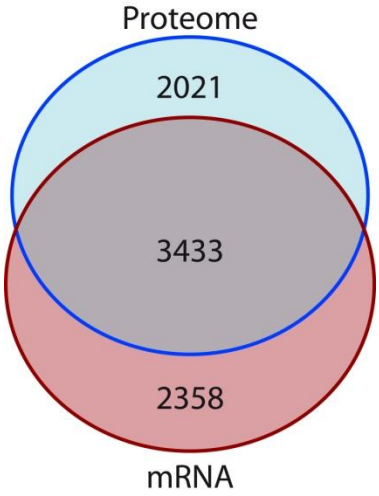


Figure 8.2 Comparison of steady state HeLa transcriptome with the quantified proteome. The microarray data for four samples of unsynchronized HeLa cells taken from the dataset of Carson *et al.* were analyzed as explained in section 8.2.10. The Venn diagram shows the overlap between two datasets based on common Entrez gene identifiers. In total 3,433 entries were common between the two datasets with a similar number of entries identified exclusively by only microarray or MS-based proteomics.

8.3.3 Analyzing proteome time course by novel bioinformatics approach

In order to discover the patterns of expressions across 6 time points we devised a new metric taking into account the maximal expression of proteins with respect to its peaking time and called it "peak time index" (Figure 8.3; Section 8.2.6). Ordering the proteins according to their "peak time index" revealed distinct up- and down-regulated clusters corresponding to each cell-cycle stage (Figure 8.4). Key players of cell cycle stages were found to show clearly discernible and already established kinetics as marked in Figure 8.4. The peak patterns were then divided into six sub-clusters (M peak, G1 peak, G1/S peak, Early S peak, Late S peak, G2 peak) depending on the concerted proteome peaking in respective cell cycle stages (annotated by blue gradient coloring in Figure 8.4). Each of these sub clusters were analyzed by a proteomic phenotypic approach using KEGG and Gene Ontology (GO) as annotational resources to reveal distinct functional characteristics and cellular contexts.

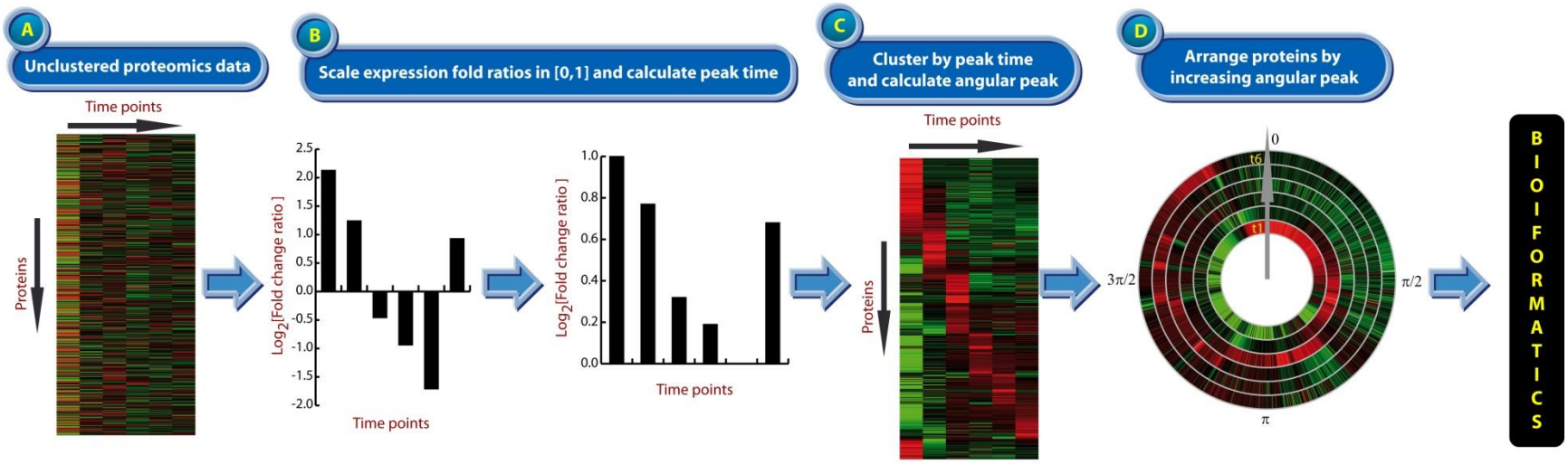


Figure 8.3 Bioinformatics workflow to cluster and circularize proteome and phosphoproteome changes. The proteome time profiles were first scaled in range [0,1] and then assigned a time peak index (section 8.2.6). The complete proteome was then clustered as per the increasing time peak index as shown in step (C).The time peak index was further transformed into an angular peak measure (section 8.2.7) in the range [0,360] degrees. This angular peak measure was used to render the *z*-transformed proteome profile as shown in step (D). Subsequent bioinformatics analysis was done using circular statistical methods on the polar coordinates of these proteins.

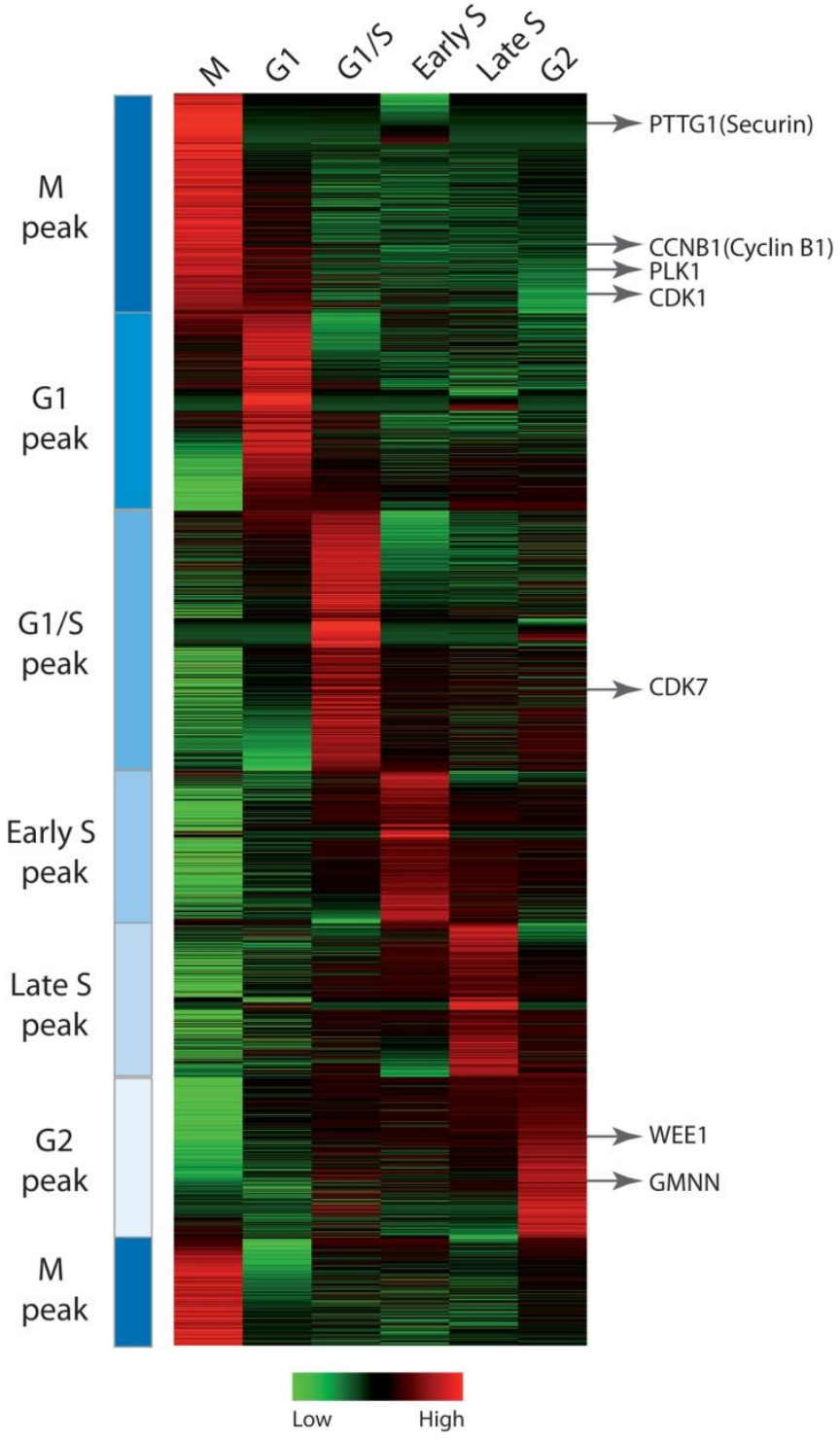


Figure 8.4 Dynamics of the proteome during the cell cycle. Proteins that were found to be regulated at least four-fold during the cell cycle were clustered in all cell-cycle stages by calculating a time peak index by weighted mean of the expression ratio of maximal expression. For each cell cycle stage, there are clear patterns of up and down-regulation.

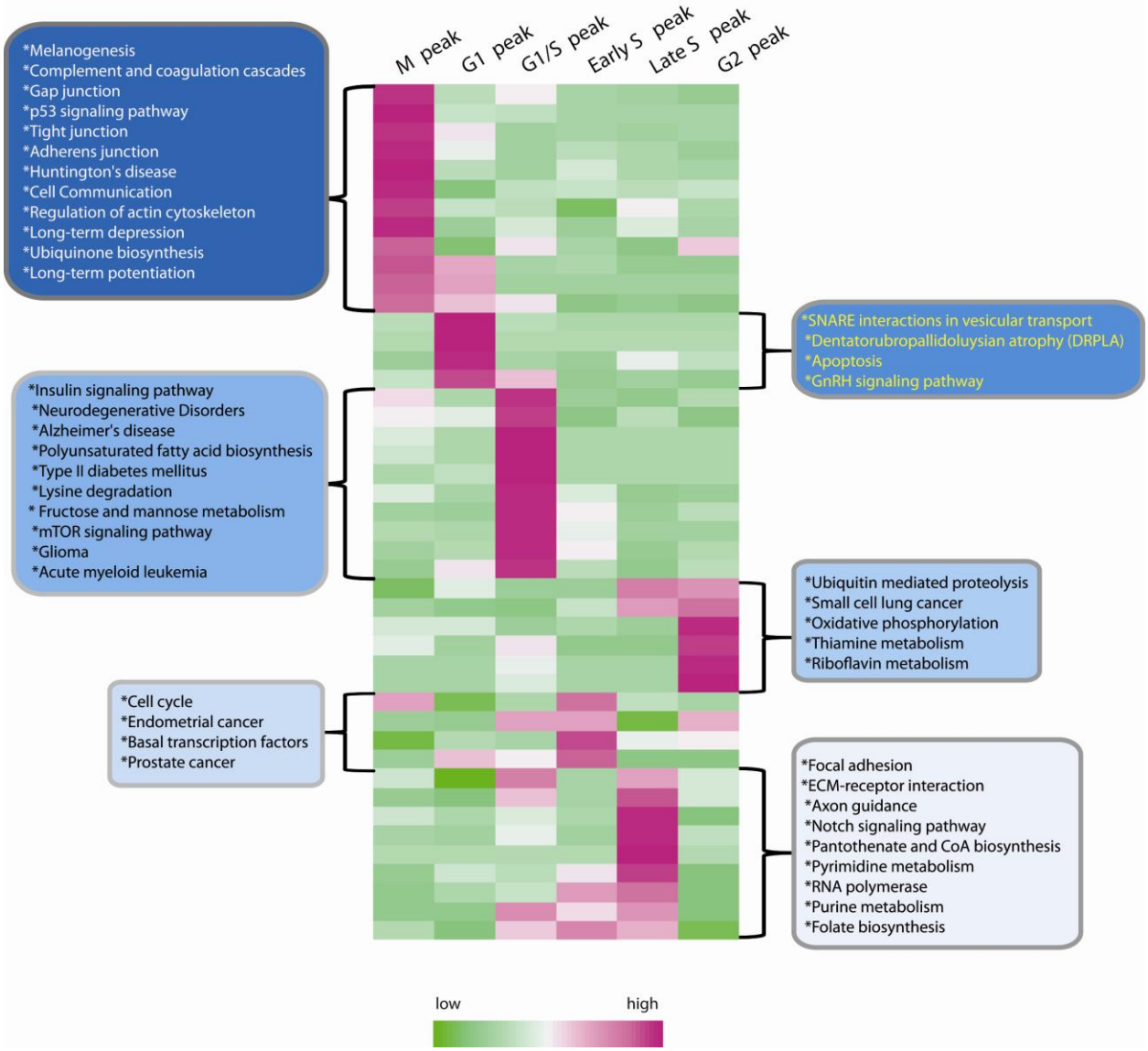


Figure 8.5 KEGG pathway enrichment analysis applied to the individual clusters derived from the cell cycle regulated HeLa proteome (marked in Figure 8.4 with blue gradient colors).

KEGG pathways enrichment based clustering provided a systems level peek into the diverse role of these proteins ranging from pivotal signaling pathways (p53, insulin, mTOR, GnRH), disease (Glioma, AML, Type II diabetes mellitus), and metabolic networks (OXPHOS) (Figure 8.5).

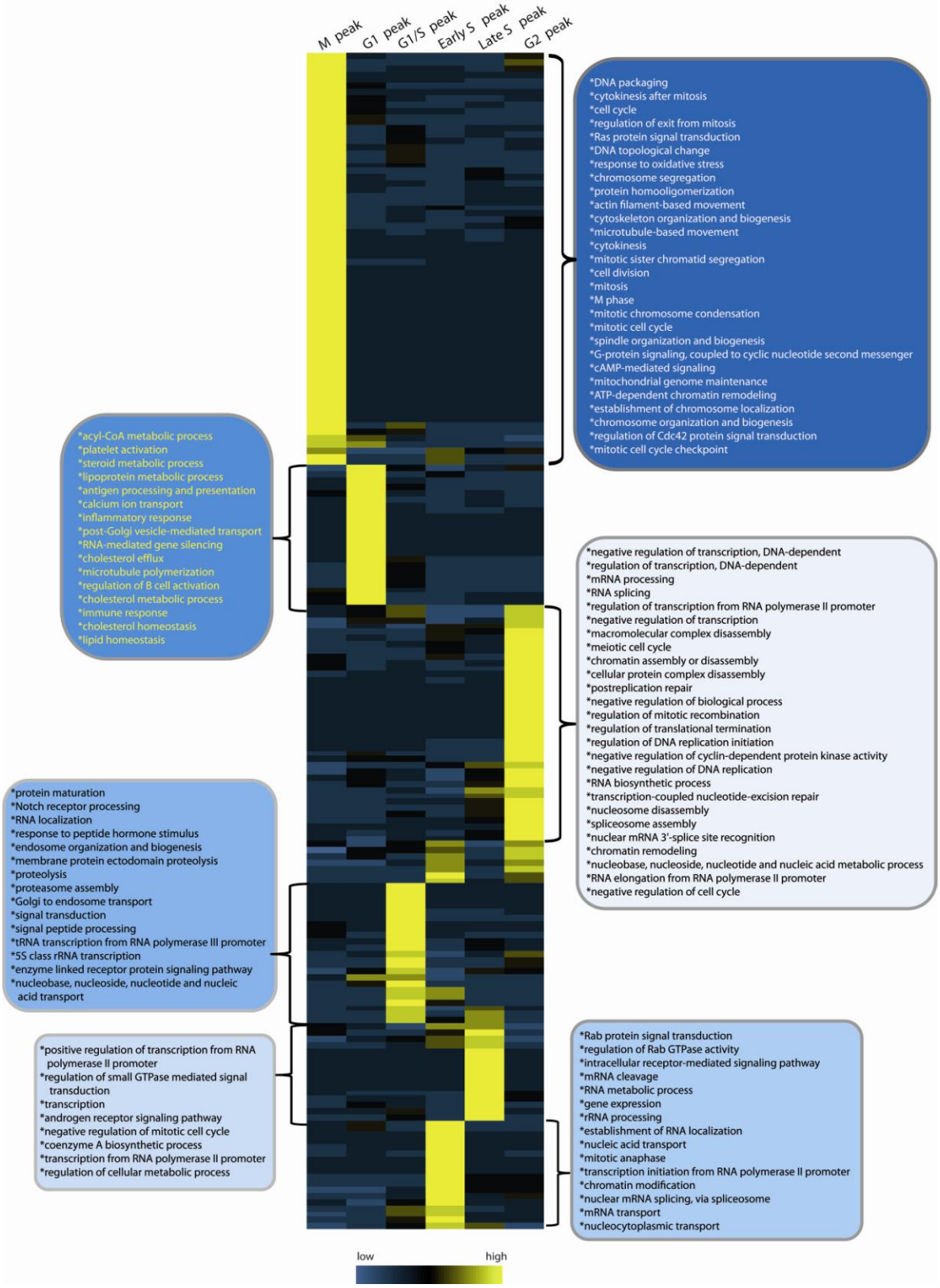


Figure 8.6 Gene Ontology (GO) enrichment analysis of biological processes (BP) applied to the individual clusters derived from the cell cycle regulated HeLa proteome (marked in Figure 8.4 with blue gradient colors)

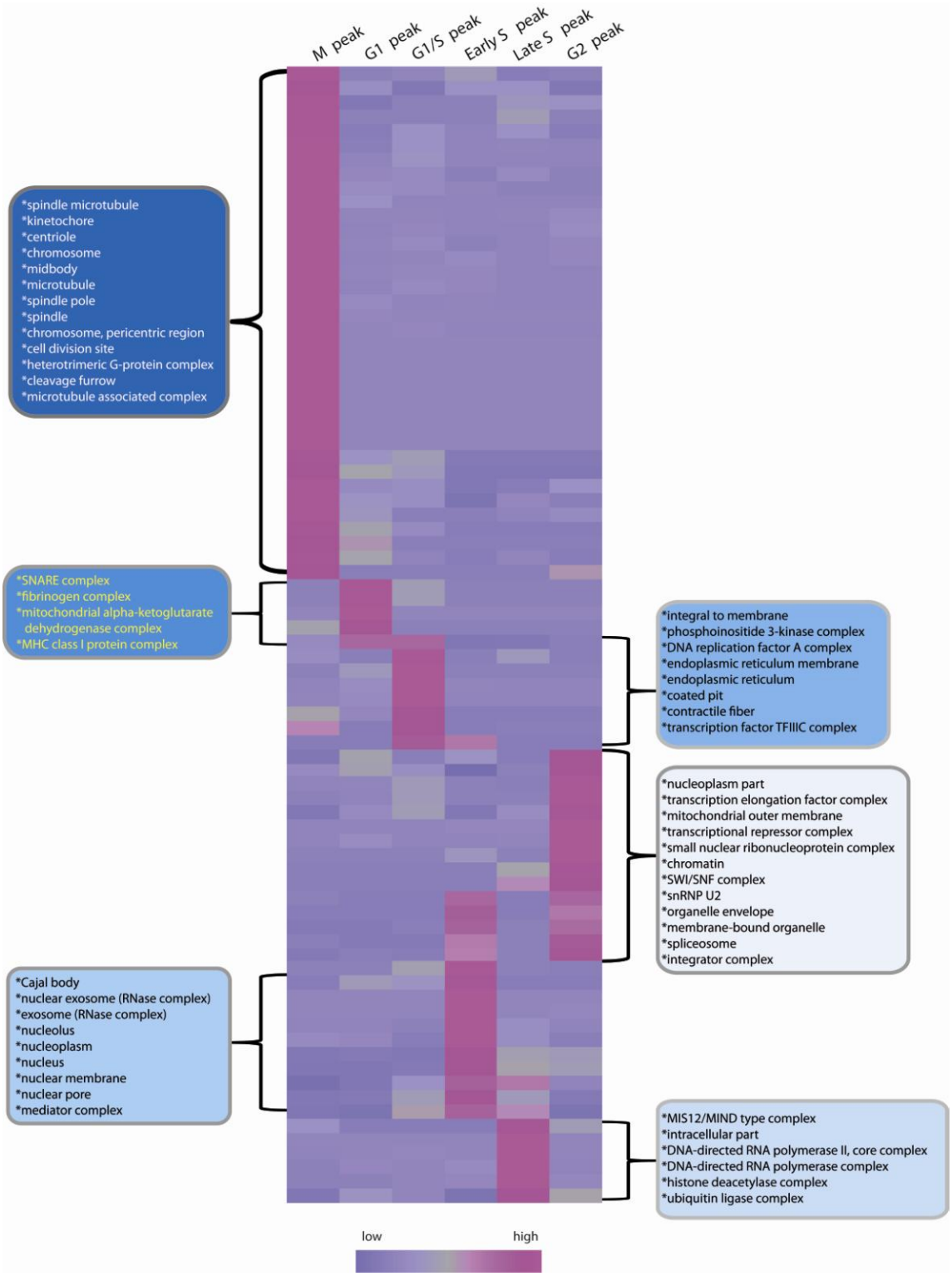


Figure 8.7 Gene Ontology (GO) enrichment analysis of cellular components (CC) applied to the individual clusters derived from the cell cycle regulated HeLa proteome (marked in Figure 8.4 with blue gradient colors)

Analysis of these clusters using GO biological process revealed clear functional enrichment of categories that would be expected in these cell cycle clusters; for instance M-phase, cell cycle, mitotic chromosome condensation among others were clearly enriched in the "M peak" cluster (Figure 8.6), and additionally many novel functions relevant to each of the six clusters shown in figure 8.4. Similarly, analysis of GO cellular components highlighted enrichment of distinct cellular compartments that are clearly defined sites of many of the biological processes and pathways found to be enriched in those clusters thereby substantiating the results found in previous analysis (Figure 8.7).

8.3.4 Directional statistics based enrichment of protein profiles reveal co-regulated complexes

To determine when in the cell-cycle specific proteins and protein groups peaked we circularized the clustered proteins on a cell-cycle timeline by mapping the peak times onto a transformed polar coordinate system[310] (Figure 8.8). Subsequently directional statistics based on the "Rayleigh test" was used to find enriched GO cellular components and complexes (see Section 8.2.7). Co-regulation of subunits within cell cycle regulated complexes namely APC/C ($P < 0.02$), Mediator complex ($P < 1E\text{-}05$), DNA replication factor A complex ($P < 0.007$) was observed (Figure 8.8). Likewise, proteins from sub-cellular organelles were found to be co-regulated, for example, mitochondrial proteins ($P < 1.6E\text{-}25$), nucleolar proteins ($P < 0.01$) and ER-Golgi components ($P < 0.001$).

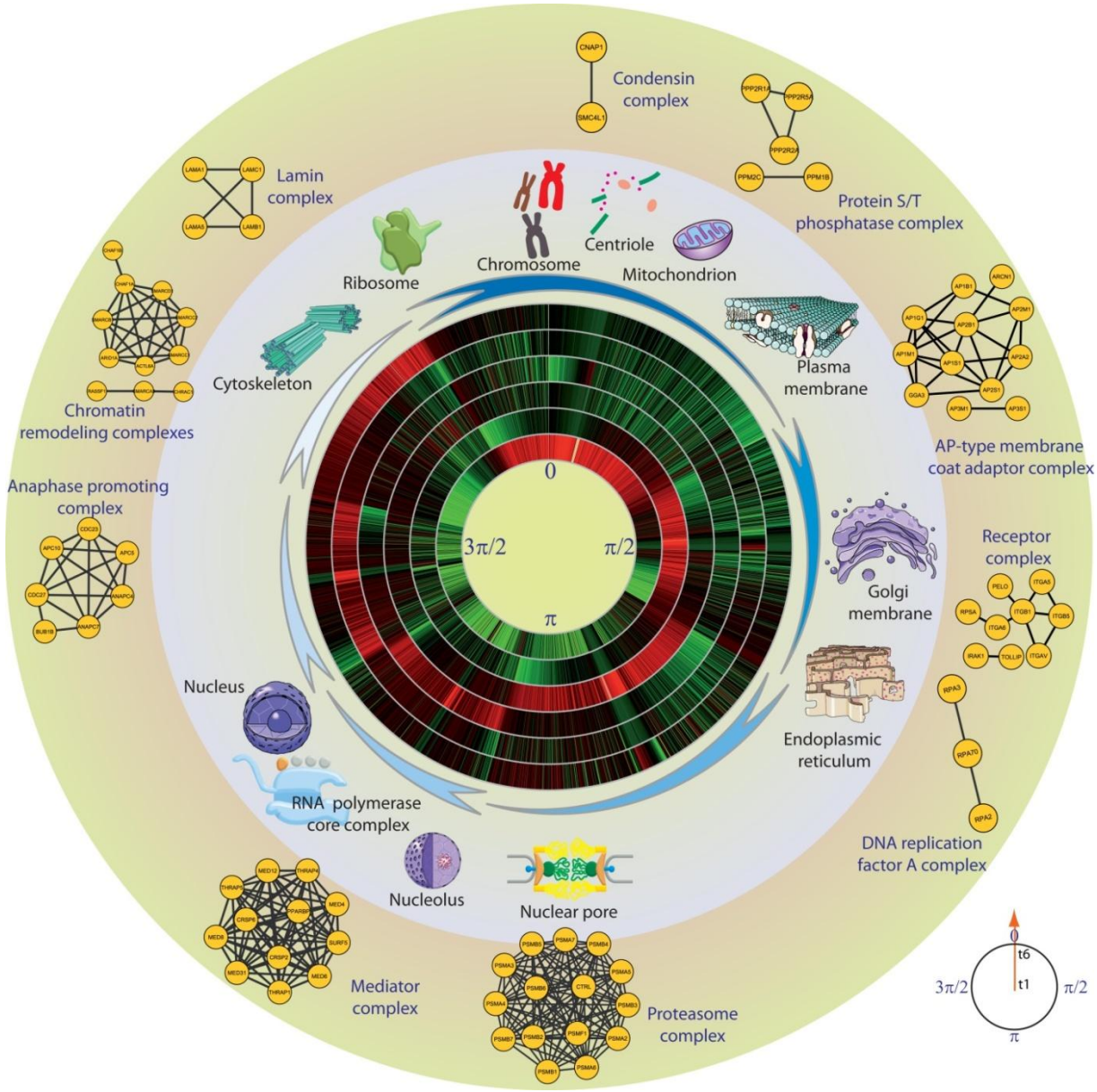


Figure 8.8 Dynamic cell cycle proteome rendered in cyclic order along with enriched gene ontology (GO) cellular compartments and complexes. The data shown in Figure 8.4 was circularized to determine the angle in the cell cycle where particular proteins peak. Around the circle co-regulated protein complexes and organellar proteins for particular cell cycle stages are indicated.

8.3.5 Proteome transcriptome comparison reveals similar depth of coverage and weak expression correlation

We next compared the dynamics of the proteome to a published cell-cycle transcriptome that made use of the same cell type and a related experimental procedure for cell synchronization[298]. Detected proteome and transcriptome overlapped to 63% and covered the expressed genome to similar depth (Figure 8.2). Steady-state message levels did not correlate well with steady state protein levels as estimated by summed peptide intensities (R=0.38, Figure 8.9A; R=0.46 when correcting for protein length, Figure 8.9B). This is not surprising since different transcripts and proteins have different half-lives.

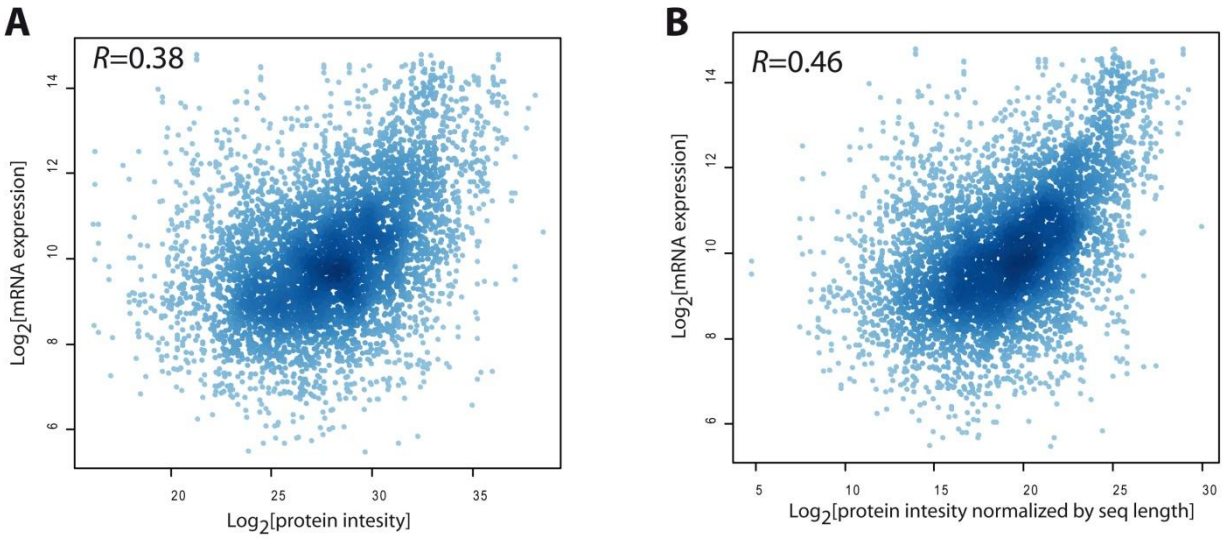


Figure 8.9 Transcriptome versus proteome comparison reveals an uncorrelated behavior (A) Summed peptide intensity vs. mRNA expression on $\log_2$ scale **(B)** Corrected summed peptide intensity vs. mRNA expression on $\log_2$ scale

However, 59% of the genes that significantly changed at the transcriptome level and that were quantified as proteins changed significantly (ratio-change >4) at the proteome level as well. We found that 21% of all observed proteins significantly change in abundance during the cell cycle, whereas only 10% of the transcriptome was reported to change[298]. This most likely reflects differences in the statistical analyses and experimental setup as well as the contribution of post-transcriptional regulation. Reassuringly, Gene Ontology (GO) analysis confirmed a cell-cycle function for proteins regulated at both the mRNA and the protein level, whereas proteins regulated at neither of the levels are preferentially involved in homeostasis and basic metabolic processes (Figure 8.10). The only functional class of proteins that are specifically regulated at the protein level but not at the mRNA level is the transcriptional machinery.

8.3.6 Analysis of cell cycle phosphorylation by ensemble bioinformatics approach

To determine phosphorylation sites that show dynamic profiles due to changes in phosphorylation state rather than due to changes in protein abundance, we normalized the measured phosphopeptide ratios by the corresponding protein changes. Subsequently we clustered the phosphoproteome time course data by increasing peak time index (section 8.2.6). The phosphoproteome is three times as large as that of our group's recent growth factor signaling study[65]. It encompasses 70% of those sites and was distributed across cellular compartments as observed before. Interestingly, the level of phosphorylation of more than half of the phosphorylation sites changed at least two-fold over the cell cycle; of these again about half were maximally phosphorylated in M-phase (Figure 8.11A). Compared to single-stimulus studies, this is a much larger proportion, highlighting the involvement of many more signaling processes in the cell cycle. Inspection of known cell-cycle-regulated phosphorylation sites showed the expected kinetics, as exemplified for the activation loop phosphorylation site (pT161) and the inhibitory sites (pT14 and pY15) of CDK1, which show opposing kinetics (Figure 8.11B).

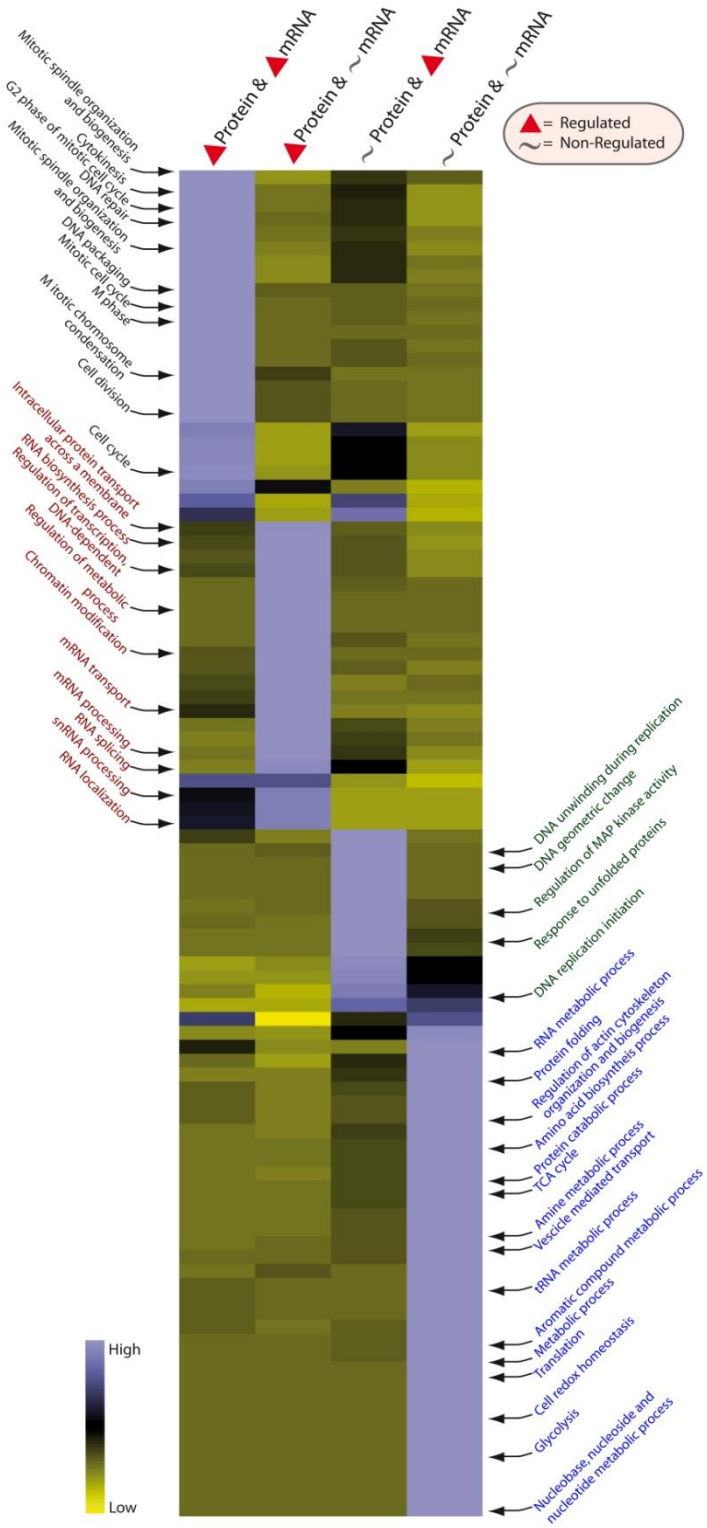


Figure 8.10 Comparison of mRNA and protein dynamics during the cell cycle. Measured protein dynamics were correlated to mRNA data published by Whitfield et al. (*3*). Proteins were grouped on the Y-axis in four categories from top to bottom: unchanging mRNA and protein, changing mRNA and unchanging protein, changing mRNA and unchanging protein and changing mRNA and changing protein. In addition, proteins were clustered on the X-axis according to their gene ontology.

8.3.7 Kinase substrate relationship prediction and novel insights into phosphorylation mediated cellular processes

Next we used NetPhosK to predict kinase-substrate relationships and constructed a heat map based on the dynamic phosphoproteome[307]. For each stage of the cell cycle this map indicates the degree of over- or under-representation of substrates of different kinase groups (Figure 8.11C).

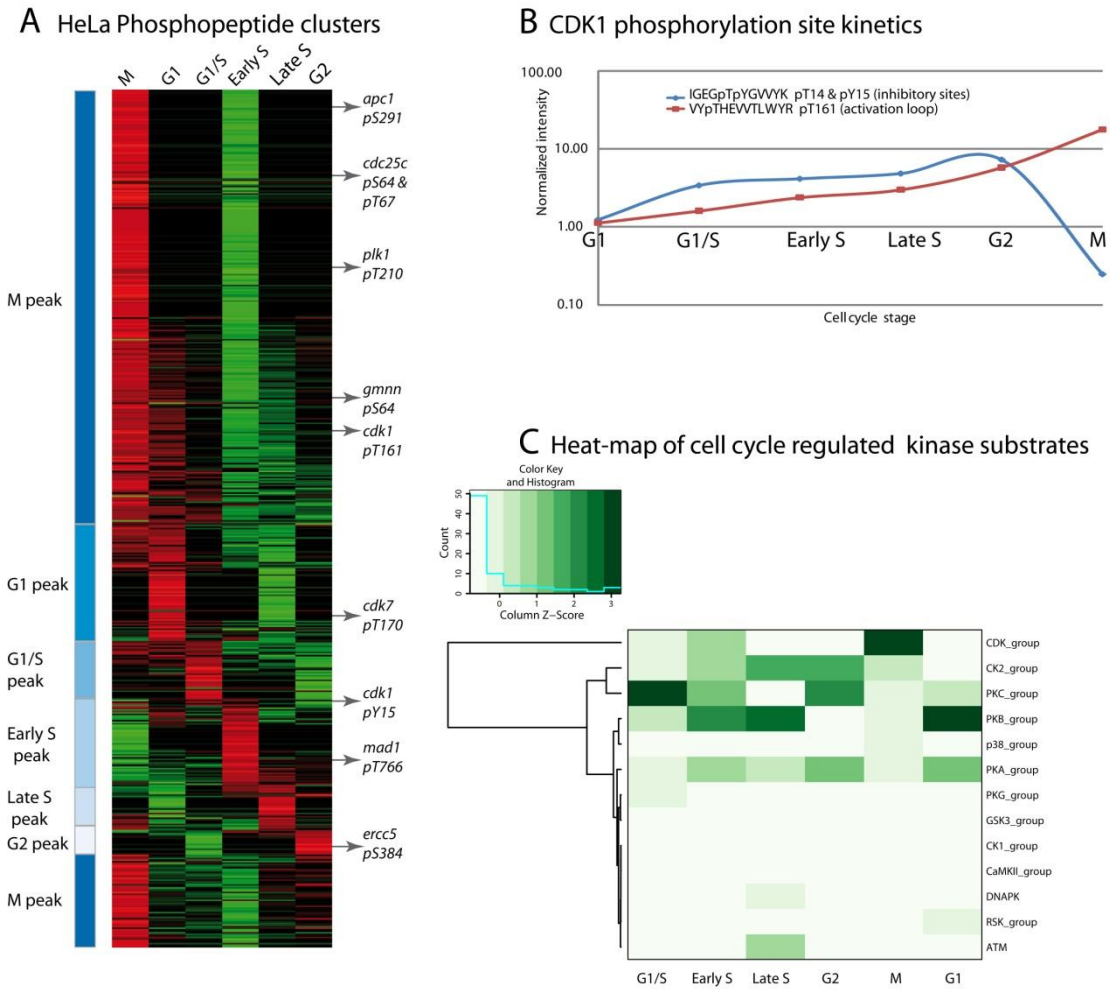


Figure 8.11 Dynamics of the phosphoproteome during the cell cycle. **A.** Clustering of regulated phosphorylation sites in all cell cycle stages as described for the proteome in Figure 8.4. More than half of all identified regulated phosphorylation sites peak in mitosis. **B.** Dynamic profile of two CDK1 phosphopeptides during the cell cycle. The activating site T161 clearly peaks in mitosis, whereas the inhibitory sites T14 and Y15 are down-regulated in mitosis. **C.** Heat map of cell-cycle-regulated kinase substrates. The NetPhosK algorithm was used to construct a heat map based on the cell cycle phosphoproteome and reveals overrepresentation of particular kinases during different stages of the cell cycle as shown. Strikingly, ATM and DNAPK kinases were found to be globally active during S phase.

As expected, predicted CDK substrates are most highly phosphorylated in M-phase. PKB/AKT-related signaling is most active in G1 whereas PKC-related substrates are preferentially phosphorylated in G1/S, which presumably reflects their growth-associated roles[311]. Interestingly, substrates of the DNA damage response (DDR) kinases ATM/ATR and DNA-PK are significantly overrepresented in S phase, which is most likely due to the coupling between DNA replication and repair. While this has been reported for individual phosphoproteins, our data show that it is a general phenomenon since 124 sites that match the ATM/ATR and DNA-PK kinase motifs peak in S phase. Many of these phosphoproteins are known to be involved in DNA repair, such as RAD50, MDC1, TP53BP1 and ERCC6. A recent phosphoproteomic study of DNA damage also identified some of these sites[312].

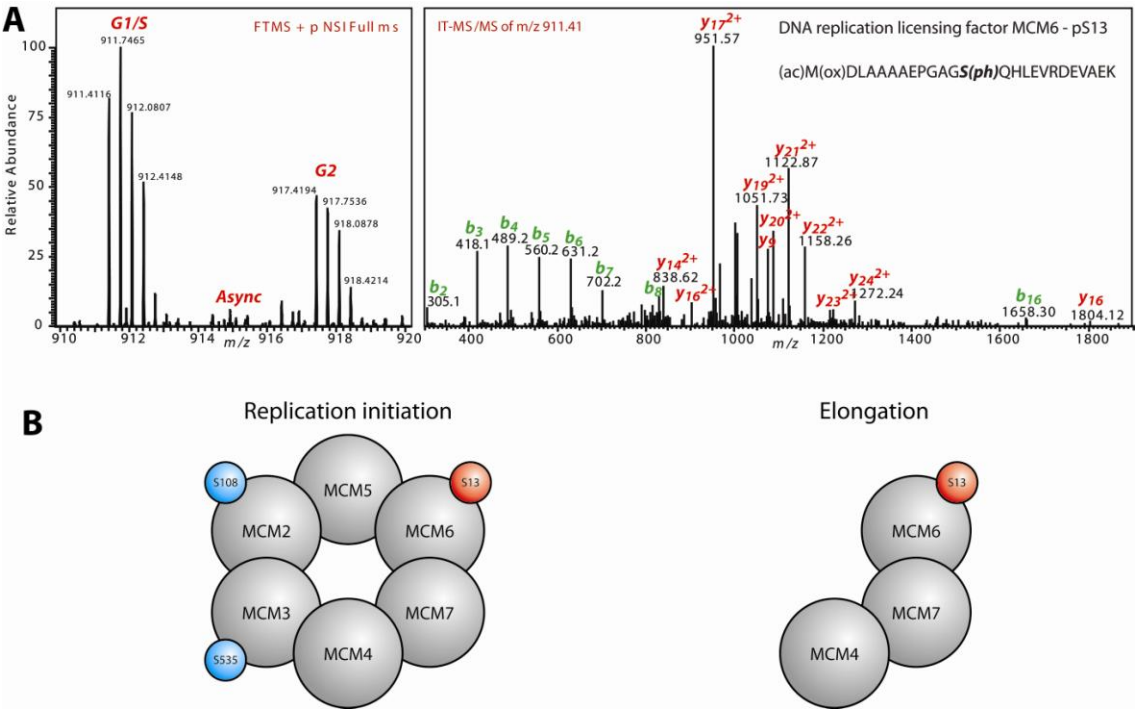


Figure 8.12 Regulation of mini-chromosome-maintenance (MCM) complexes in response to DNA damage. (A) Tandem mass spectrum of a phosphopeptide derived from MCM6 that contain the pS13. **(B)** Left panel, MCM2-7 is one of several complexes involved in initiation of DNA replication. In response to DNA damage, the subunits MCM2 and MCM3 are phosphorylated on S108 and S535, respectively. We identify a novel pS-Q phosphorylation site (S13) on the MCM6 subunit, which is phosphorylated during S-phase. Right panel, during the elongation phase of DNA replication, only the subcomplex MCM4/6/7 is part of the replication fork. The novel phosphorylation site on MCM6 provides a plausible mechanism by which the DDR kinases can halt ongoing DNA replication in response to DNA damage.

The DDR kinases[313] coordinate cell-cycle checkpoints and DNA repair mechanisms and almost exclusively phosphorylate substrates with linear sequence motifs that match pS/T-Q[314]. Out of 479 pS/T-Q sites in the data set we selected 21 sites in proteins that are regulated at multiple levels. These were analyzed with the NetworKIN algorithm, which increases the accuracy in classifying kinase-substrate relationships by combining prediction of kinase substrate motifs with contextual information[308]. This analysis tied 14 substrates to ATM and DNA-PK, of which five are known to be part of the DDR apparatus. The DDR kinases also phosphorylate the MCM2 and MCM3 subunits of the hexameric mini-chromosome-maintenance (MCM) complex[315], which is essential for initiation of DNA replication. In agreement with this, we show that S108 of MCM2 is most highly phosphorylated during S-phase. However, this does not explain how the DDR kinases can halt ongoing DNA replication in response to DNA damage, since only the MCM4/6/7 helicase subcomplex is involved in the elongation process[316]. We identify a novel pS-Q site within this subcomplex (S13 of MCM6, Figure 8.12A) which is phosphorylated specifically during S-phase and hence provides a plausible mechanism by which DDR kinases could inhibit ongoing DNA replication (Figure 8.12B).

8.3.8 Systematic study of cell cycle control regulation by integrating proteome, phosphoproteome and transcriptome

Apart from regulated phosphorylation, targeted protein degradation is a key regulatory mechanism in cell-cycle control. To analyze our data with respect to degradation motifs, we integrated the different levels of signal processing and gene expression at the protein and transcriptome levels. Degradation signals are significantly overrepresented in proteins that are cell-cycle regulated; this is true both for transcriptional regulation (KEN boxes, $P < 10^{-18}$; PEST regions, $P < 0.002$), regulation of protein levels (KEN boxes, $P < 0.002$; PEST regions, $P < 0.02$) and periodic phosphorylation (KEN boxes, $P < 10^{-4}$; PEST regions, $P < 0.001$). As expected, proteins that are regulated at multiple levels are more enriched in degradation signals than the individual sets; for example, proteins that are regulated by phosphorylation and transcription show a 2.6-fold enrichment for KEN boxes, whereas phosphorylation alone leads to a 1.4-fold enrichment only.

If gene products that are regulated at the phosphoproteome or proteome level are highly enriched for proteins that have important functions during the cell cycle, then knock-down of these proteins should result in cell-cycle defects. Indeed, of 39 genes that show a cell-cycle phenotype in a siRNA pilot study[317], 14 were regulated at the phosphoproteome in M-phase, and 29 at the proteome level (Table 8.1).

Gene name	RNAi phenotype observed in study by Neumann *et al.* [Nature Methods 2006. 3(5) 385-390]	Timing of phenotype	Proteomics Max Difference	M-phase specific Phosphorylation?
RGPD5	Medium mitosis and apoptosis phenotype	Early onset	2.23	-
NU153	NUP153 - Medium mitosis and apoptosis phenotype	Early onset	1.55	-
SYNE2	Medium mitosis, shape and apoptosis phenotype	Early onset	1.5	-
RAD21	Medium mitosis and apoptosis phenotype	Medium onset	1.76	-
ANC1	medium all	Early mitosis	1.96	-
NUP62	Medium shape	Late	1.64	-
SEH1L	Medium mitosis and apoptosis phenotype	Early onset	1.37	-
CDC27	Weak mitosis and medium apoptosis phenotype	Early onset	1.38	-
ANC2	Medium all	Late onset	1.93	-
NUP37	Medium shape	Late	1.23	-
NU107	NUP107 - Medium mitosis & apoptosis phenotype	Early onset	1.15	-
MD2L1	Medium shape	Late onset	1.38	-
NUMA1	NUMA1 - Medium mitosis and apoptosis phenotype	Early onset	1.01	-
ANC5	medium all	Early mitosis	1.19	Yes
LMNA	Weak mitosis and medium apoptosis phenotype	Late onset	0.95	Yes
SUV42	Weak mitosis and medium apoptosis phenotype	Late onset	1.45	Yes
APC10	Weak mitosis and medium apoptosis phenotype	Late onset	1.16	Yes
POM121	Medium mitosis and apoptosis phenotype	Early onset	0.99	Yes
CBX1	Weak mitosis and medium apoptosis phenotype	Medium onset	0.16	Yes
CBX3	Medium mitosis and apoptosis phenotype	Medium onset	0.13	Yes
DUS3	medium shape,	early onset	0.82	-
COPB	Strong apoptosis	Early	0.17	Yes
CAPG	Medium mitosis and shape	Medium	0.25	Yes
SEC13	Medium all	early	0.27	Yes
MP2K3	Medium shape	Late	1.5	-
CBX5	Weak mitosis and medium apoptosis phenotype	Medium onset	0.16	Yes
CDK7	Medium shape	Medium onset	0.75	-
EGFR	Weak mitosis and medium apoptosis phenotype	phospho?	0.37	Yes
H2AY	Weak mitosis and medium apoptosis phenotype	Medium onset	0.85	-
CENPH	Weak mitosis and medium apoptosis phenotype	Medium/Late	0.61	Yes
CENPB	Medium mitosis and apoptosis phenotype	Late onset	0.7	Yes
KIF11	Strong mitosis phenotype	Early	1.82	-
KIF23	Medium all, stronger mitosis	Early	2.3	-
AURKB	Medium mitosis, shape and apoptosis phenotype	Early onset	1.69	-
DUS14	Medium shape, weak mitosis and apoptosis	Early	3.28	-
TPX2	Medium mitosis, shape and apoptosis phenotype	Early onset	2.83	-
SMC4	Weak mitosis and medium apoptosis phenotype	Early onset	3.11	-
PLK1	Strong mitosis phenotype	Early	4.48	-
CCNB2	Weak mitosis and medium apoptosis phenotype	Late onset	3.05	-

Table 8.1 Comparison of our cell cycle regulated proteome and phospoproteome to RNAi phenotype study by Neumann *et al.*[317]

We found 27 gene products that exhibit periodic regulation of phosphorylation states, protein levels, and transcription (Table 8.2).

HUGO	IPI id	Phosphorylation sites	Degradation signal
DSP	IPI00013933	40(9)	D/K/P
BIRC2	IPI00013418	8(1)	-
GMNN	IPI00026309	9(1)	K/P
UNG	IPI00011069	6(2)	P
TROAP	IPI00029680	8(1)	P
BRIP1	IPI00012500	10(1)	D/P
NUSAP1	IPI00000398	15(4)	K
KIF23	IPI00293884	17(1)	K/P
ZNF24	IPI00306446	5(5)	P
LMNB1	IPI00217975	22(4)	P
GPSM2	IPI00642575	8(1)	D
RAD18	IPI00024579	6(2)	K/P
TMPO	IPI00030131	14(1)	D/P
GAS2L3	IPI00185219	5(2)	P
MKI67	IPI00413173	3(3)	P
ATAD2	IPI00170548	11(2)	-
NUP35	IPI00329650	26(12)	-
C4A	IPI00032258	5(4)	D/P
PLK1	IPI00021248	3(1)	-
RRM2	IPI00011118	5(1)	K
SHCBP1	IPI00168691	6(1)	P
CASP2	IPI00291570	2(2)	-
PRR11	IPI00305822	8(3)	D/K/P
TACC3	IPI00002135	12(1)	K/P
SFRS12	IPI00375462	4(1)	-
TPX2	IPI00008477	10(1)	D/K
KPNA2	IPI00002214	16(7)	-

Table 8.2 Protein regulated at multiple levels during the cell cycle. The table contains the HGNC name, the IPI identifier, the number of detected phosphorylation sites (with the number of periodic phosphorylation sites in parenthesis) and the presence/absence of degradation motifs (D for D-box, K for KEN box and P for PEST region) for each of the 27 proteins.

Among these are two genes, LMNB1 (Lamin B1) and TMPO (Lamina-associated polypeptide 2), which encode lamina-related proteins that influence nuclear envelope stability. TMPO contains a Lamin-B-binding domain that mediates its interaction with LMNB1[318]. This interaction is

disrupted by phosphorylation of TMPO in mitosis[319]. Our data now identifies the M-phase-specific site of phosphorylation (S306), which is located within the Lamin-B-binding domain of TMPO, as being responsible for this disruption. Other lamina-related gene products, LMNA (Lamin A), LMNB2 (Lamin B2), and LBR (Lamin B Receptor) were also found to cycle at multiple levels. Very recently, it was shown that the regions that surround lamina-associated chromosomal domains contain binding sites for certain transcription factors[320]. Most of these are E2Fs and other G1/S transcription factors, which is intriguing since TMPO has been shown to regulated cell-cycle progression via the Rb–E2F pathway[321]. Thus it is tempting to speculate that lamina-related proteins regulate G1/S transcription both by interacting with known G1/S transcription factors and by binding close to their downstream target genes.

8.4 Discussion

In this work we present the first global and unbiased analysis of proteome and phosphoproteome dynamics during the cell cycle at a depth of about 6,000 proteins and the quantitation of more than 18,000 unique phosphorylation sites. Our quantitative proteomics dataset provides a valuable resource for large-scale studies of in vivo phosphorylation dynamics at a systems-biology level. Complemented with novel bioinformatics analysis and the inferences gathered therein we provide a global compendium of cell cycle related pathways, functions and components. We expect it to be useful for the cell-cycle and cancer communities as it directly connects gene expression changes with protein regulatory information at a proteome-wide level.

9. Protein localization assignment in brown and white adipose tissue mitochondria by multiplexed quantitative proteomics and systematic bioinformatics approach

The bioinformatics approach for mitochondrial protein localization discussed in this project is part of the work included in a manuscript under submission:

Francesca Forner, **Chanchal Kumar**, Christian A. Luber, Martin Klingenspor and Matthias Mann

Pathway analysis of mitochondria in brown versus white adipocytes by quantitative proteomics

9.1 Introduction

The imminent goal of understanding a cell at the 'systems level' hinges on the accurate knowledge of the dynamic, spatial and temporal aspects of its sub-cellular components including RNAs, proteins and metabolites. One definitive step towards this is mapping the sub-cellular localization of proteins which provides key insights into their cellular function and interaction with other entities[62]. Until now systematic experimental studies of subcellular localization of proteins have been performed with the help of cellular fractionation or fluorescent microscopy. The first proteome wide *localizome* study of budding yeast using green fluorescent protein (GFP) fusion proteins has been published[322]. In more complex eukaryotes and mammalian systems GFP based methods face numerous challenges, and has been recently shown to result in experimental artifacts such as translocation to nucleus causing spurious localization[323]. Alternatively, antibodies or other affinity reagents have been successfully used to visualize protein sub-cellular localization. While this method has several advantages over GFP based methods, one of its largest drawbacks is the lack of availability of comprehensive antibody inventory to probe the protein localization on a global and high-throughput scale. One of the most comprehensive studies based on fluorescently labeled antibodies coupled with confocal microscopy has recently reported the localization of 1,899 human gene products[324].

Additionally, numerous bioinformatics methods for protein localization prediction have been in use for many years and have provided pointers towards function of hypothetical or novel and uncharacterized proteins[325-327]. As they are mainly based on sequence features they can be applied for genome wide studies of protein localization predictions for organisms which have already been sequenced. But because of the underlying algorithms and the input constraints, they are very specific for particular class of proteins (eukaryotic vs. prokaryotic, secretory, mitochondrial, nuclear, chloroplast) and therefore limited or specialized in application[328-331]. Moreover, they provide mere indications towards protein localizations which need to be ultimately verified by cell-biological or biochemical methods.

Proteomics methods specifically applied to characterization of sub-cellular organelle proteomes has been referred in literature as *Organellar Proteomics* and refers to the ensemble of innovative experimental methods, MS techniques and specialized bioinformatics algorithms employed for characterization of organelle proteomes[332,333]. Organellar proteomics methods have been successfully applied to unravel the proteomes of various organelles including mitochondria, nucleolus, isolated synaptic vesicles, clathrin coated vesicles, endosomes, phagosomes, endoplasmic reticulum, and Golgi apparatus, as well as Golgi-derived COPI vesicles[58,333-336]. Quantitative proteomics methods add another dimension to organellar proteomics thereby facilitating study of the dynamics of organelles and their constitutive proteomes under various cellular states[45]. SILAC is a valuable method for quantitative proteomics and has found numerous applications in basic and translational research. Recently organellar proteomics methods based on SILAC have been successful employed to map the dynamics of protein trafficking in human nucleolus and to model the phagosome maturation process[337,338]

As protein localization is contingent upon many factors which are at play in a given cellular context, a dynamic picture of protein localization calls for a special approach and methodology. Following this idea we devised a comprehensive framework and workflow for assigning simultaneous protein localization in two sub-cellular organelles, compartments and fractions by integrating SILAC, organelle enrichment, quantitative mass-spectrometry analysis and a novel probabilistic bioinformatics approach. Given the limitations of obtaining totally purified organelles and the fact that in any fractionation method a part of the protein population is always

present as contaminant in different sub-fractions, we reasoned that a probabilistic approach of localization assignment in separate fractions is more rational. The workflow described is suited for simultaneous protein localization assignment in two sub-cellular compartments or fractions and could easily be extended to more fractions.

9.2 Materials and Methods

9.2.1 Preparation of SILAC reference

3T3-L1 and brown preadipocytes were sub-cultured and differentiated in DMEM supplemented with 10% dialyzed fetal bovine serum (Gibco) and antibiotics in 5% CO2 at 37°C. SILAC labeling was performed as described[339] with L-Lysine-13C6,-15N2 and L-Arginine-13C6,-15N4. 3T3-L1 preadipocytes were grown and differentiated as described previously (Kratchmarova et al. 2002). Brown preadipocytes were obtained from C.R. Kahn's laboratory and differentiated as described[340]. Cells were harvested with Trypsin (Gibco), diluted with DMEM supplemented with protease inhibitors (Roche) and centrifuged at 1000 g for 10 min. Cells were then resuspended with 250 mM sucrose, 10 mM Hepes pH 7.4, 0.1 mM EGTA supplemented with protease inhibitors (Roche) and washed twice. The suspension was homogenized on a 7 ml Dounce homogenizer. Mitochondria and nuclei were isolated as described below. The crude mitochondrial fraction was purified on a 30% Percoll self-forming gradient. Equal amounts of brown and white adipocyte mitochondria were mixed 1:1 based on the protein amount and served as the internal standard (am-IS). The post mitochondrial fraction (PMF) was the supernatant obtained after the first centrifugation at 10000 g (see below). The PMF was concentrated by with 5000 MWCO membranes (Millipore).

9.2.2 Preparation of mitochondrial sample

Interscapular brown adipose tissue and epididymal white adipose tissue from C57BL/6 mice (5-10 weeks) were excised, immersed in HBSS, cleaned free of connective tissue under a binocular microscope, minced and digested with 1 mg/ml collagenase A (Roche) at 37°C for 30 min. After digestion, the slurry was passed through 250 μm mesh opening fiber material (Sefar) and centrifuged at 500 g for 1 min. The floating adipocytes were removed with a plastic pipette and

centrifuged three additional times in HBSS. No visible stromal vascular fraction was present after the fourth centrifugation. Floating adipocytes were homogenized using a 10 µm clearance cell homogenizer (Isobiotec) pre-cooled in ice. Membrane disruption was checked with Trypan blue staining. The tissue homogenate was centrifuged twice at 800 g for 10 min at 4°C to pellet the crude nuclear fraction. The supernatant was centrifuged at 10000 g for 10 min at 4°C. The crude mitochondrial pellet was resuspended in 250 mM sucrose, 10 mM Hepes pH 7.4, 0.1 mM EGTA supplemented with protease inhibitors (Roche). The suspension was further centrifuged at 7000 g for 10 min at 4°C and purified with the protease treatment as described[341]. For the localization study, mitochondria from the brown and from the white adipose tissues were mixed 1:1 with post mitochondrial (PMF) and nuclear fractions (N) isolated from the SILAC-labeled brown adipocytes or 3T3-L1 respectively (see section 9.2.1). For quantitative proteomics, mitochondria from the brown and from the white adipose tissues were each mixed 1:1 with the am-IS based on protein amount (Bradford).

9.2.3 Protein fractionation and mass spectrometric analysis of proteins and relative quantitation.

Protein samples were separated on 1D gels, trypsin-digested and extracted as described[342]. Peptides were desalted and concentrated on C_{18} stage tips as described[235] and analyzed by LC-MS/MS on a LTQ-Orbitrap mass spectrometer (Thermo Fischer Scientific) connected to an Agilent 1200 nanoflow HPLC system via a nanoelectrospray source (Proxeon Biosystems). MS full scans were acquired in the orbitrap analyzer by using internal lock mass recalibration in real-time. MS full scans were acquired in two different m/z ranges (350-1000 and 1000-1800). Tandem mass spectra of the 6 most intense ions of the lower mass range and of the 4 most intense ions of the higher mass range were simultaneously recorded in the linear ion trap. Peptides were identified from MS/MS spectra by searching them against the IPI mouse database (version 3.24) using the Mascot search algorithm (www.matrixscience.com) and SILAC pairs were quantified by MaxQuant[47].

9.2.4 Finite Mixture modeling and Bayesian approach for protein localization

The section explains the Bayesian inference methodology of localization probability assignment for brown adipocytes experiment shown in Figure 9.1A. The same approach was applied to the 3T3-L1 white adipocytes experimental setup. Below we discuss the two cases namely case 1: mitochondrial (*Mito*) vs. post mitochondrial fraction (*PMF)*, and case 2: mitochondrial (*Mito*) vs. nuclear fraction (*Nuc*).

Case 1: We used the statistical method of finite mixture modeling to model the abundances of the $\log_2\left(\frac{PMF}{Mito}\right)$ ratios based on the assumption that the "*Mito*" and "*PMF*" protein ratios originate from two different Gaussian distributions $f(x;\mu_{Mito},\sigma_{Mito})$ and $f(x;\mu_{PMF},\sigma_{PMF})$ respectively. To estimate these distributions we used expectation maximization (EM) algorithm[343] with an initial assignment of the protein ratios $x=\log_2\left(\frac{PMF}{Mito}\right)$ to either the mitochondrial class (C_{Mito}) or the PMF class(C_{PMF}) by the criteria: $\begin{Bmatrix} if\,(x \le -1.5), x \in C_{Mito} \\ if\,(x > -1.5), x \in C_{PMF} \end{Bmatrix}$. The cutoff -1.5(on $\log_2$ scale) was chosen as it corresponds to ~ 3 fold up/down ratios and enables a preliminary separation of the dataset as observed from manual inspection. The EM algorithm iteratively finds the best estimates for the Gaussian distributions for the two classes given the SILAC ratios for all the proteins identified in that experiment. The class membership (or localization) probability for each protein was calculated by the equations defined below. Briefly, given a protein ratio $x=\log_2\left(\frac{PMF}{Mito}\right)$, the following two equations calculate the class membership probability P_{Mito} *and* P_{PMF} of its membership to mitochondrial class (C_{Mito}) and PMF class(C_{PMF}) respectively.

$$P_{Mito} = P(C_{Mito} \mid Ratio = x) = \frac{P(Ratio = x \mid C_{Mito}) * P(C_{Mito})}{P(Ratio = x \mid C_{Mito}) * P(C_{Mito}) + P(Ratio = x \mid C_{PMF}) * P(C_{PMF})}$$

And,

$$P_{PMF} = P(C_{PMF} \mid Ratio = x) = \frac{P(Ratio = x \mid C_{PMF}) * P(C_{PMF})}{P(Ratio = x \mid C_{Mito}) * P(C_{Mito}) + P(Ratio = x \mid C_{PMF}) * P(C_{PMF})}$$

Where,

$$P(Ratio = x \mid C_{Mito}) = f(x;\mu_{Mito},\sigma_{Mito})$$

$$P(Ratio = x \mid C_{PMF}) = f(x;\mu_{PMF},\sigma_{PMF})$$

Case 2: As in case 1 expectation maximization (EM) algorithm was used to model the abundances of the $\log_2\left(\frac{Nuc}{Mito}\right)$ ratios based on the assumption that the "*Mito*" and "*Nuc*" protein ratios originate from two different Gaussian distributions $f(x;\mu_{Mito},\sigma_{Mito})$ and $f(x;\mu_{Nuc},\sigma_{Nuc})$ respectively. The initial assignment of the protein ratios $x=\log_2\left(\frac{Nuc}{Mito}\right)$ to either mitochondrial class (C_{Mito}) or nuclear class(C_{Nuc}) was done by the criteria $\begin{Bmatrix} if\ (x \le -1.5), x \in C_{Mito} \\ if\ (x > -1.5), x \in C_{Nuc} \end{Bmatrix}$ similar to case 1**.** The EM algorithm iteratively finds the best estimates for the Gaussian distributions for the two classes given the SILAC ratios for all the proteins identified in that experiment. These estimates for the two class distributions were in turn used to assign the class membership probabilities. Given a ratio, $x=\log_2\left(\frac{Nuc}{Mito}\right)$**,** the following two equations calculate the membership probability P_{Mito} *and* P_{Nuc} of its membership to mitochondrial class (C_{Mito}) and nuclear class(C_{Nuc}) respectively.

$$P_{Mito} = P(C_{Mito} \mid Ratio = x) = \frac{P(Ratio = x \mid C_{Mito}) * P(C_{Mito})}{P(Ratio = x \mid C_{Mito}) * P(C_{Mito}) + P(Ratio = x \mid C_{Nuc}) * P(C_{Nuc})}$$

And,

$$P_{Nuc} = P(C_{Nuc} \mid Ratio = x) = \frac{P(Ratio = x \mid C_{Nuc}) * P(C_{Nuc})}{P(Ratio = x \mid C_{Mito}) * P(C_{Mito}) + P(Ratio = x \mid C_{Nuc}) * P(C_{Nuc})}$$

Where,

$$P(Ratio = x \mid C_{Mito}) = f(x;\mu_{Mito},\sigma_{Mito})$$

$$P(Ratio = x \mid C_{Nuc}) = f(x;\mu_{Nuc},\sigma_{Nuc})$$

This complete analysis was done using "MCLUST" version 3 package[344] in the R statistical environment[345].

9.2.5 Categorization of proteins in discrete classes based on probability cutoff

The proteins were finally assigned to separate categories based on the localization probability values. Below we discuss the two cases pertaining to our study of brown adipocytes, namely case 1: mitochondrial (*Mito*) vs. post mitochondrial fraction (*PMF)*, and case 2: mitochondrial (*Mito*) vs. nuclear fraction (*Nuc*). A similar approach was employed for classifying proteins in the 3T3-L1 white adipocyte experimental setup

Case 1: Based on the P_{Mito} *and* P_{PMF} membership probabilities for each protein, we categorized the proteins in 3 exclusive categories by the following criterion: (a) MITO: if (P_{Mito} >= 0.75), (b) BORDER: if((P_{Mito} < 0.75) AND (P_{PMF} < 0.75)), and (c) Non-MITO: if (P_{PMF} >= 0.75).

Case 2: Based on the P_{Mito} *and* P_{Nuc} membership probabilities for each protein, we categorized the proteins in 3 exclusive categories by the following criterion: (a) MITO: if (P_{Mito} >= 0.75), (b) BORDER: if((P_{Mito} < 0.75) AND (P_{Nuc} < 0.75)), and (c) Non-MITO: if (P_{Nuc} >= 0.75).

9.2.6 Gene Ontology based localization concordance matrices

The accuracy metrics for our localization approach were defined with respect to known GO cellular compartment annotations available for IPI version 3.24 downloaded from the EBI GOA website (version `gene_association.goa_mouse.32.gz`). As the premise of our approach was to identify mitochondrial proteins we specifically choose the available annotation pertaining to "mitochondrion" (GO:0005739) for each of the four dataset(two each for BAT and WAT). The percentage of true-positive and false-negative mitochondrial proteins assigned by this approach were quantified by two metrics namely True Localisation Percentage (TLP) and False Localization Percentage (FLP) as defined below:

$$\text{TLP} = \frac{(\text{Proteins in MITO class}) \text{ AND } (\text{Annotated as Mitochondrial (GO:0005739)})}{\text{All Proteins Annotated as Mitochondrial(GO:0005739) in dataset}} * 100$$

$$\text{FLP} = \frac{(\text{Proteins in Non-MITO class}) \text{ AND } (\text{Annotated as Mitochondrial (GO:0005739)})}{\text{All Proteins Annotated as Mitochondrial(GO:0005739) in dataset}} * 100$$

9.3 Results

9.3.1 Multiplexed proteomics approach to obtain mitochondrial localizations in Brown and White Adipose Tissue

Mitochondria (Mito) from brown adipose tissue (BAT) and white adipose tissue (WAT) were purified by density gradient centrifugation. In separate experiments nuclear (Nuc) and the post-mitochondrial fractions (PMF) were isolated from SILAC-labelled brown adipocytes and 3T3-L1. We then mixed BAT mitochondria with (1) SILAC-labeled brown adipocyte PMF and (2) with SILAC-labeled brown adipocyte Nuc in a 1:1 ratio (Figure 9.1A). Likewise, we mixed WAT mitochondria with (1) SILAC-labeled 3T3-L1 PMF and (2) with SILAC-labeled 3T3-L1 Nuc in a 1:1 ratio (Figure 9.1B). The four samples were separated in 40 fractions on 1D gels and measured by high resolution mass spectrometry.

The PMF/Mito and Nuc/Mito protein ratios provided the relative protein abundance over the three subcellular fractions, based on the assumption that relatively high PMF/Mito or Nuc/Mito ratios designated non-mitochondrial proteins (higher abundance in the PMF or Nuc fraction) whereas relatively low PMF/Mito or Nuc/Mito ratios designated mitochondria-associated proteins. Plotted ratio distributions were more distinctly bimodal for brown adipocytes/BAT than for 3T3-L1/WAT. Manual inspection of the datasets confirmed that the left tails (lower ratios) were highly enriched in mitochondrial proteins, as expected from the lower PMF/Mito or Nuc/Mito ratios. In total 3,689 proteins were identified with quantified SILAC ratios in at least one of the four experiments.

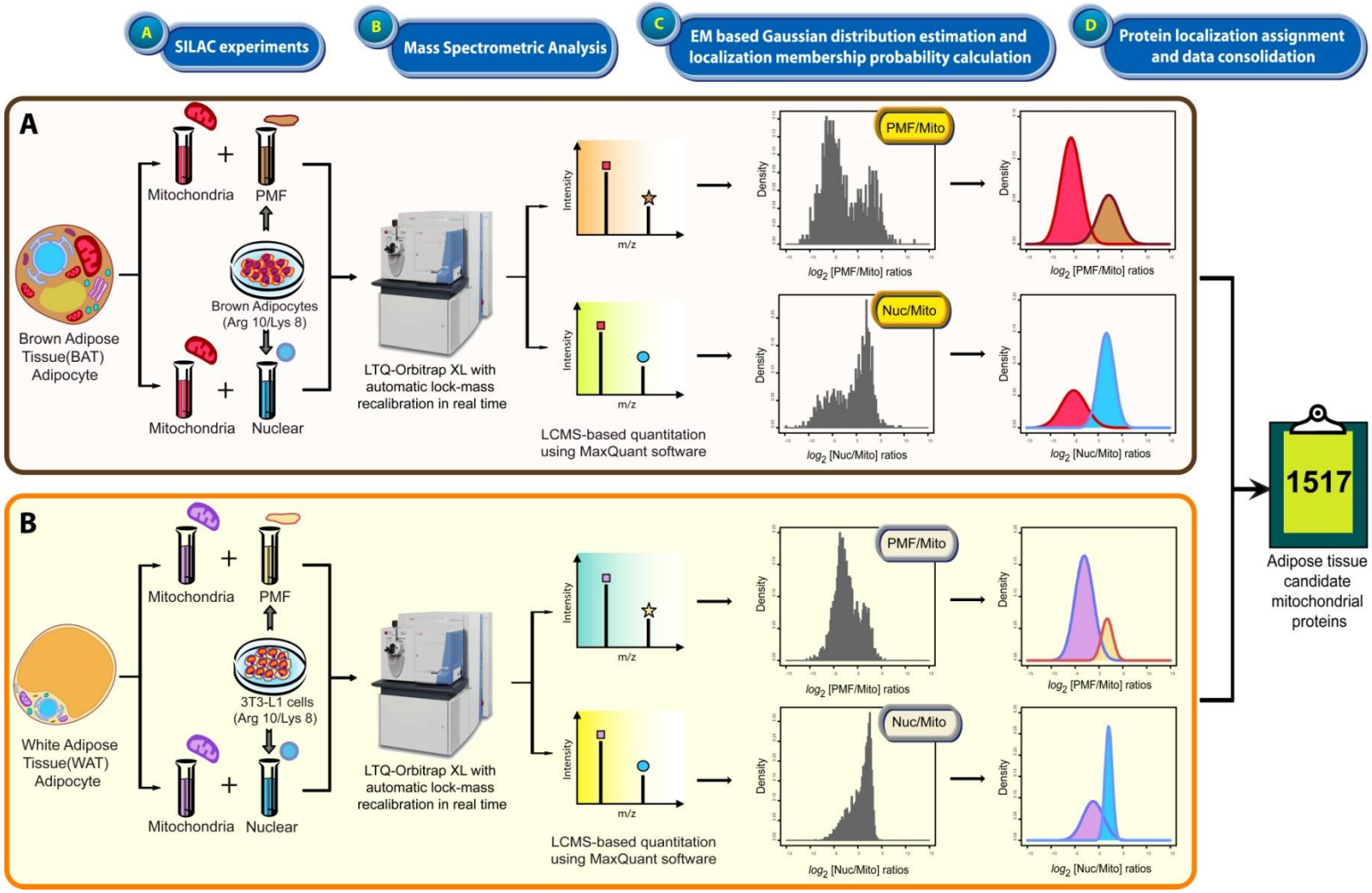


Figure 9.1 Workflow of mitochondrial protein localization based on quantitative proteomics and bioinformatics analysis (A). Mitochondria were enriched from the brown fat and mixed on one to one ratio with nuclear and post-mitochondrial enriched fractions obtained from SILAC labeled cell line of brown adipocytes. Plotted protein ratios distributions appeared bimodal and were interpolated with the expectation maximization algorithm (EM) to calculate the probability of mitochondrial localization. (B) The procedure described in (A) was repeated with mitochondria enriched from white adipocytes, whereas nuclear and post-mitochondrial fractions were enriched from SILAC-labeled 3T3-L1.

9.3.2 Probability based localization assignment of mitochondrial proteins

Our next goal was to establish core mitochondrial proteomes from this rich dataset. One of the simplest empirical methods could have been to choose an arbitrary cut-off of PMF/Mito and Nuc/Mito ratios and then categorise the proteins as mitochondrial, PMF or Nuclear. This criterion is very inconsistent for multiple datasets of similar origin but with distinct features like the one here, as in each case we would have to define a separate cut-off. Moreover this approach is not amenable to automation and depends on manual curation, validation, individual expertise and bias. We wanted to avoid all of these above pitfalls in our data analysis and sought to devise a generic bioinformatics approach which could be consistently applied to any such dataset. Given the data in hand and the nature of the problem to be solved we adopted a probabilistic model based on Bayesian inference to assign probabilities to the proteins based on their SILAC ratios. Bayesian methods have found myriad applications in bioinformatics and have been successfully applied to sequence analysis, microarray data analysis, network inference and systems biology[346]. Our proposed analysis methodology is a two step process. In the first step we model the abundances of the protein ratios based on the assumption that the ratios would originate from two different Gaussian distribution functions pertaining to (a) Mitochondrial proteins, and (b) the PMF or Nuclear proteins, under the given experimental strategy. Expectation Maximization (EM) algorithm[343] was employed to model these Gaussian distributions. In the second step we use these estimates to assign a probabilistic measure of locality (P_{Mito}, $P_{\text{PMF}}/P_{\text{Nuc}}$) using sets of Bayesian equations (for details see section 9.2.4).

9.3.3 Grouping of proteins in organelle classes based on Bayesian probabilities

For each experiment we categorized the proteins as “MITO”, “BORDER” and “Non-MITO” based on conservative probability values (see section 9.2.5 and Figure 9.2). The rationale behind defining a BORDER class was twofold - on a computational level these are proteins which could not be confidently assigned to either of the two categories based on their obtained probability values and so they are borderline cases, and from a biological perspective these could be proteins which are present on the interface of the two organelles (like proteins associated with the outer membrane) or proteins which have more tightly coupled dynamics between the organelles. Moreover, GO annotations for these BORDER proteins revealed that they contained some interesting mitochondrial proteins in each case. Finally, we applied a very conservative threshold

of $P_{Mito} > 0.75$ to define core mitochondrial proteome in each experiment. Though the BORDER proteins were not included in the core set, in subsequent discussion we show the implication of their inclusion or exclusion on our results assessment metrics. Figure 9.2 provides the summary of results obtained by our categorization approach for the two experiments against the backdrop of their quantitative proteomics ratios (PMF/M or N/M) and the estimated Gaussian distributions. The number of proteins in each category is represented with the class labels. It is evident that setting an arbitrary cut-off would have been difficult and completely arbitrary to categorize the data. Interestingly, in each set of experiments the combined set of MITO and Border class together contained (1) 97% and 94% of GO annotated mitochondrial proteins for WAT and (2) 95% and 97% of GO annotated mitochondrial proteins for BAT respectively; thereby demonstrating the power of our method in sorting out true mitochondrial proteins from non-mitochondrial ones. As a first step towards validation we checked for the classification of known mitochondrial residents (e.g. respiratory chain subunits, Krebs cycle enzymes, fatty acid oxidation enzymes, translocases of the outer membrane) and cytoplasmic/mitochondrial isoforms of a certain protein (e.g. aspartate aminotransferase) were correctly classified.

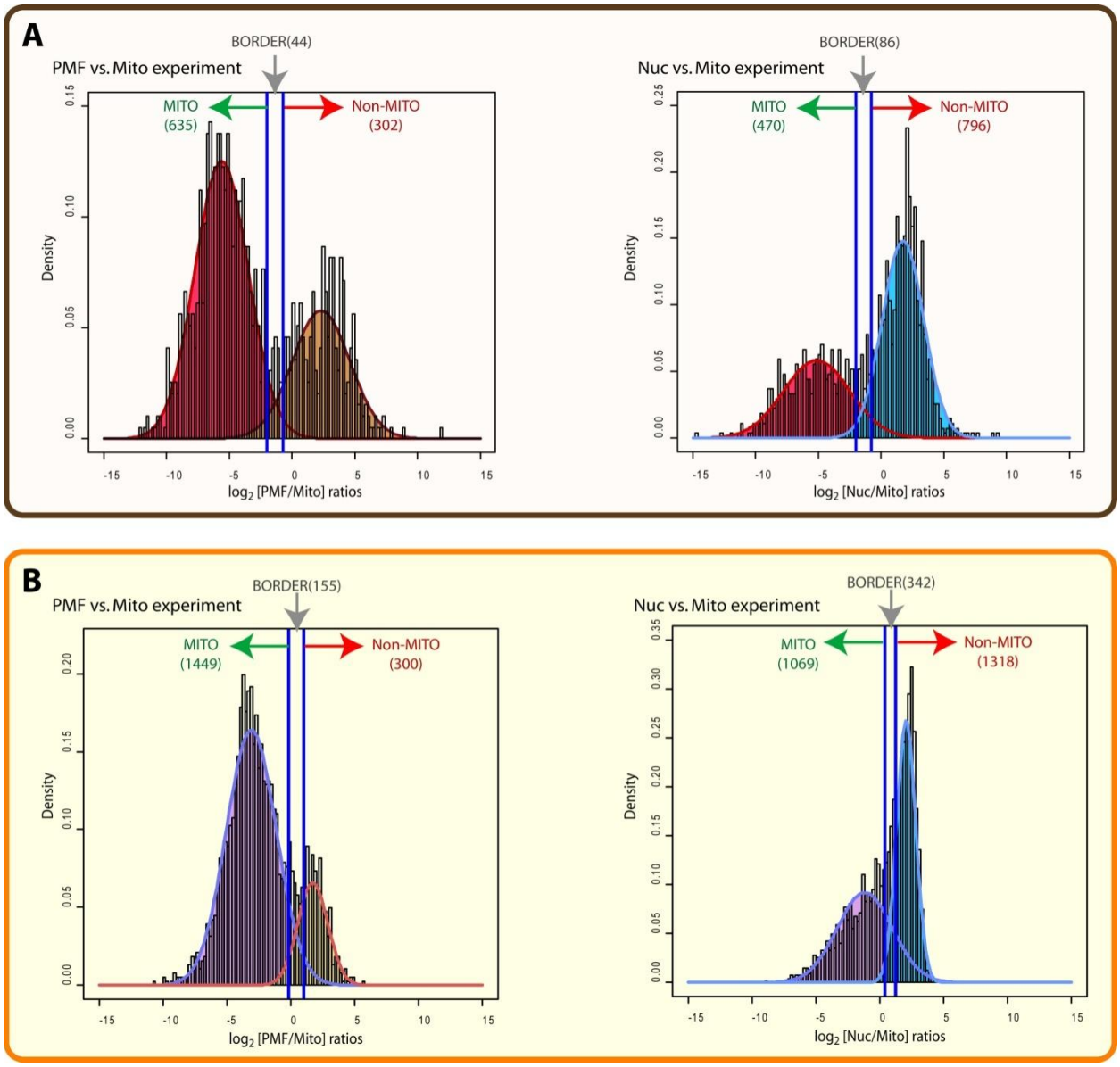


Figure 9.2: Categorization of proteins in three classes (MITO, Border, and Non-MITO) based on the calculated localization probability values. **(A)** The number of proteins categorized in BAT experiments against the backdrop of quantitative proteomics ratios (in black bars) and the estimated Gaussian distributions (enveloped by colored areas). In each plot the left distribution pertains to estimated mitochondrial population. The ratio cutoffs corresponding to the categorization criterion are shown as vertical blue lines and the number of proteins in each class is provided in parentheses. **(B)** The number of proteins categorized in WAT experiments. The details of the legends are as in (A)

9.3.4 Concordance of probabilistic localization with Gene Ontology annotations

To provide a quantitative measure to the accuracy of our method we defined two metrics. True Localisation Percentage (TLP) and False Localization Percentage (FLP) which would be respectively analogous to *true-positive* and *false-negative* rates in machine learning vocabulary but defined with respect to Gene Ontology(GO) annotations[48] (see section 9.2.6 for definition). GO annotations have been used as the benchmark in many studies on localization predictions, therefore we decided to use GO as the reference for calculating the (true/false) localization percentages for each dataset[214,347]. We compared our localization assignments with the GO cellular compartment annotations available for IPI Mouse 3.24 database from GOA website (Figure 9.3).

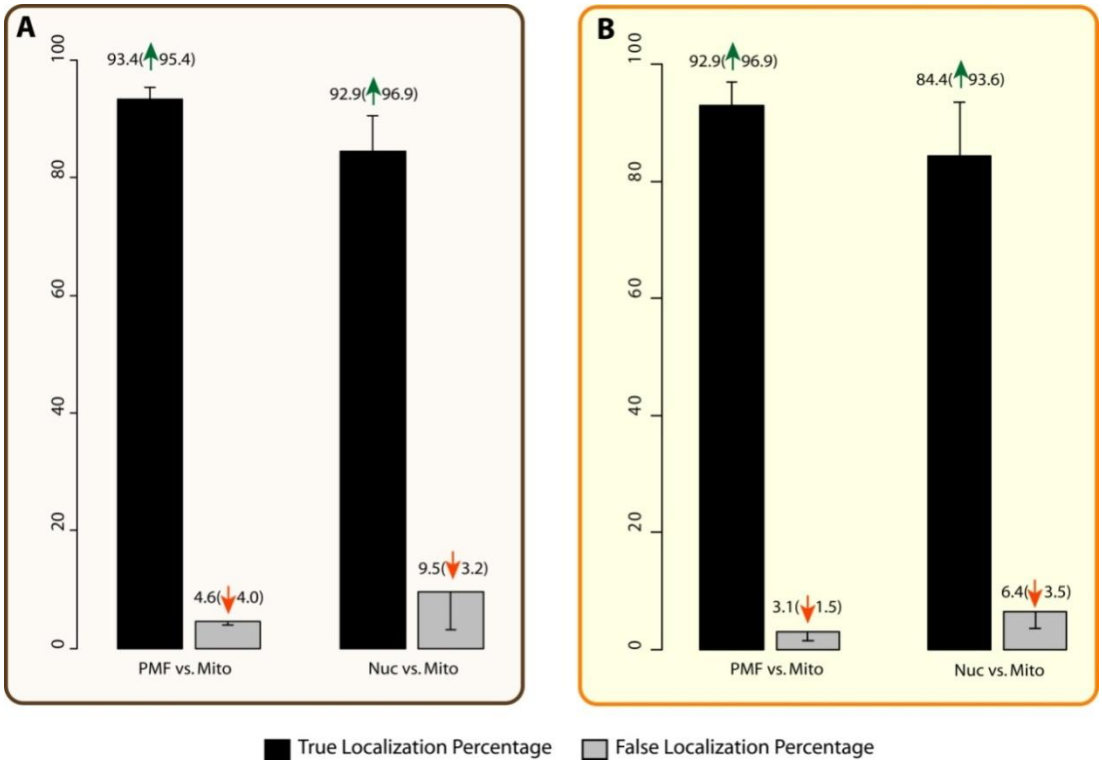


Figure 9.3 Localization concordance results for the mitochondrial localization experiments (A) Localization concordance results for the BAT mitochondrial localization. The black bars show the True Localization Percentage (TLP) and the grey bars represent False Localization Percentages (FLP) for MITO class proteomes (P_{Mito}>= 0.75). The error bars TLP show the gain % in TLP when Border proteins are included in the MITO set. The number on top of the green bars denotes the TLP for the core proteome sets with the increased TLP in parentheses (with a green arrow). The error bars on FLP shows the decrease % in FLP when multiply localized mitochondrial proteins are removed from FLP calculations. The number on top of the red bars denotes the FLP for the core proteome sets with the decreased FLP in parentheses (with a red arrow). **(B)** Localization concordance results for the WAT mitochondrial localization. The details of legends are as in (A)

9.3.4.1 High accuracy of mitochondrial localization in brown adipose tissue (BAT)

Figure 9.3A shows the results for TLP and FLP of the core mitochondrial proteomes for the two BAT experiments. For the PMF vs. Mito experiment we get TLP of 93% and FLP of 5% while for the Nuc vs. Mito experiment we get TLP of 93% and FLP of 9%. The relatively lesser TLP and relatively higher FLP for the Mito vs. Nuc experiments (as compared to PMF vs. Mito ones) may be due to the fact that there exists a strong trafficking of proteins between nucleus and mitochondria as suggested by large body of literature on mitochondrial-nuclear communication[348]. This is further substantiated by our observation on the increase in TLP when we included GO annotated mitochondrial BORDER proteins in the core proteome set. In the PMF vs. Mito experiments the TLP increased to 95 % (an increase of 2%) but in the Nuc vs. Mito exp the increase was 97 % (an increase of 5%). This higher gain in the latter case could be a reflection of the dynamics of nuclear-encoded mitochondria-targeted proteins in the particular cellular contexts during our experimentation. As the highest FLP in this case was 9% and though it was comparable w.r.t the experimental and computational localization assignment methods, but still we sought to probe the reason behind this. After careful study of the GO annotations for the proteins which accounted for FLP calculation we observed that most of them indeed had multiple localizations besides mitochondria, so they were not exactly "false-negatives" in the true sense and could be localised to other compartments. Therefore to get a more logical statistic we recalculated the FLP by removing proteins which had any other GO localization that could be part of PMF(peroxisome, membrane, lysosome, ER, cytoplasm) or Nuclear fractions (nucleus,nucleolus), in addition to mitochondrial annotations. Thereby the corrected-FLP was now 4%(a decrease of 1%) for the BAT(PMF/Mito) and 3%(a decrease of 6%) for BAT(Nuc/Mito) (Figure 9.3 A). In total we could confidently assign mitochondrial localization to 650 proteins across these 2 experiments.

9.3.4.2 High accuracy of mitochondrial localization in white adipose tissue (WAT)

Figure 9.3B shows the results for TLP and FLP of the core mitochondrial proteomes for the two WAT experiments. For the PMF vs. Mito experiment we get TLP of 93% and FLP of 3% while for the Nuc vs. Mito experiment we get TLP of 84% and FLP of 6%. As in the case of BAT the relatively lesser TLP and relatively higher FLP for the Mito vs. Nuc experiments (as compared to PMF vs. Mito ones) may be due to the fact that there exists a strong trafficking of proteins

between nucleus and mitochondria. Similarly, we observed a remarkable increase in TLP when we included GO annotated mitochondrial BORDER proteins in the core proteome set. In the PMF vs. Mito experiments the TLP increased to 97 % (an increase of 4%) but in the Nuc vs. Mito exp the increase was 94 % (an increase of 10%). This was again similar to the results in case of BAT localizations. Again, similar to the BAT case, we recalculated the FLP by removing proteins which had any other GO localization that could be part of PMF (peroxisome, membrane, lysosome, ER, cytoplasm) or Nuclear fractions (nucleus,nucleolus), in addition to mitochondrial annotations. Thereby the corrected-FLP was now 1.5% (a decrease of 1.5%) for the WAT(PMF/Mito) and 5%(a decrease of 1%) for WAT(Nuc/Mito) (Figure 9.3B). In total we could confidently assign mitochondrial localization to 1,512 proteins across these two experiments.

9.3.5 Integration of multilevel sub-cellular localization information to elucidate mitochondrial proteome of mouse adipose tissue

The two parallel localization experiments for each of the adipose tissue types i.e. BAT and WAT (Figure 9.1) provides us with two levels of confidence for a protein to be localized in either mitochondrion versus its post mitochondrial fraction or nuclei. Subsequently, as described in section 9.3.3 and illustrated in figure 9.2, for each experiment we categorize each protein as either being MITO, BORDER or Non-MITO by using conservative probability cutoffs (section 9.2.5). Furthermore, we used this dual localization evidences for each tissue type to define the core mitochondrial proteome of that tissue type. The classification contingency matrix along with the numbers of proteins identified as mitochondrial or non-mitochondrial are illustrated in Figure 9.4. In BAT we putatively assign 650 proteins to mitochondria (out of 1,753) and in WAT we assign 1,512 proteins to mitochondria (out of 3,345). Though in terms of the numbers the BAT mitochondrial proteome is approximately one-third of the WAT mitochondrial proteome, this disparity diminishes when we look at the percentage coverage of BAT and WAT mitochondrial proteome w.r.t the quantified proteome in each tissue type. BAT mitochondrial proteins cover roughly 37% of the total tissue specific organelle proteome while WAT covers 45%. Lastly, we use the union of these two tissue type mitochondrial proteomes to arrive at 1,517 putative mitochondrial proteins in adipose tissue of mouse.

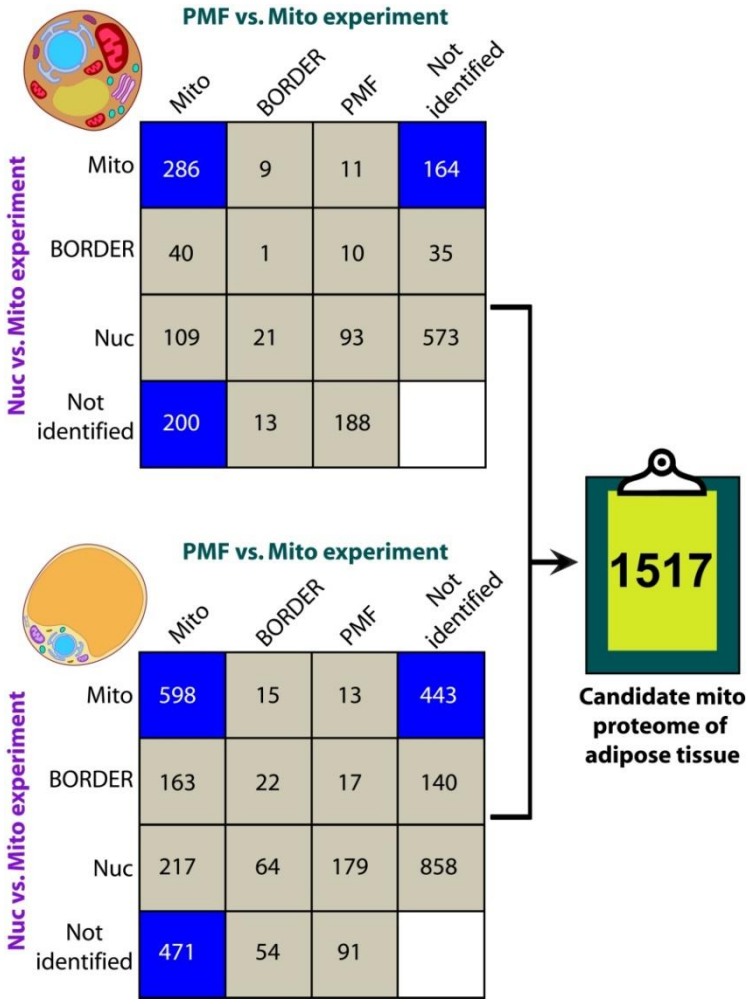


Figure 9.4 contingency matrixes for combining evidence from two parallel experiments each in BAT and WAT to define core-mitochondrial proteome of adipose tissue (WAT). The class assigned (Mito, Border, Nuc/PMF) in each experiment (Nuc vs. Mito and PMF vs. Mito; Figure 9.1) were combined to assign a final class to the proteins as shown in the matrix. Blue boxes contain the mitochondrial protein numbers in BAT (upper matrix) and WAT (lower matrix). In total we obtained 1,517 core mitochondrial proteins.

9.4 Discussion

Protein localization studies in sub-cellular organelles have become a mainstay in current systems biology enterprises, and are definitely the first steps towards realizing their promises and goals[147,349]. In recent years many proteomics methods based on MS technology have been reported for establishing the protein localization and organelle *localizomes*[332-334]. The most

effective of these methods combine elaborate cellular fractionation, mass-spectrometry and quantitative proteomics. Protein Correlation Profiling (PCP) and Localization of Organelle Proteins by Isotope Tagging (LOPIT) are the representative approaches in this category[104,350]. While they provide definite results they also require simultaneous analysis of many fractions and are limited by the efficiency of fractionation, most importantly they require prior knowledge of organelle marker proteins. We devised a simpler and generic approach whereby we build upon the strengths of quantitative proteomics and SILAC and do away with multiple fractionations. Furthermore we use a probabilistic localization method with simplest computational assumptions (or model) and do not require prior knowledge of organelle marker proteins. As a proof of concept we use this framework to assign putative localization to proteins identified from BAT and WAT mitochondria.

Mitochondria have been one of the most widely studied sub-cellular organelles and they play pivotal roles in myriad biological processes including growth, division, energy metabolism and apoptosis[351-353]. Due to their expansive role in cellular processes they are also implicated in myriad diseases including obesity, diabetes, cancer, neurodegenerative and cardio-vascular disorders[354,355]. The advent of high throughput "omics" disciplines has helped tremendously in generating parts list of mitochondria at various levels of organization including genome, proteome, metabolome and interactome; thereby providing valuable insights into its biology and cellular function[356]. The study of mammalian mitochondrial proteomes is a challenge owing to its dynamic constitution, and the complexity of the cellular milieu in these systems. In the recent past proteomics and functional genomics approaches have been employed for establishing the mitochondrial proteome of mouse and humans[57,59,342,357]. According to the latest estimates there are ~1500 mitochondrial proteins in humans[358]. While recent systemic analysis approaches are trying to enumerate the complete mitochondrial proteome in model organisms[359], there is still clear consensus that many more mitochondrial proteins remain to be discovered and characterized[360]. Moreover, study of the tissue specific organellar proteome is of immense value as it correctly portrays the physiology which is often not captured by studying cell lines of similar origin[342]. Therefore we sought to explore the mitochondrial proteome from brown and white adipose tissue of mouse as a model system for humans[361].

In two sets of parallel experiments we were able to quantify and elucidate 650 (BAT) and 1,512 (WAT) putative mitochondrial proteins according to very high probability scores obtained by our bioinformatics approach. These proteins show very high concordance with the known Gene Ontology mitochondrial localizations. Furthermore, by combining the localization evidences from these two separate tissue specific experiments we were able to derive a set of 1,517 putative mitochondrial proteins in mouse adipose tissue. This candidate mitochondrial catalogue was further used in a setup for systems level analysis of *in vivo* quantitative profiling of mitochondrial proteome of BAT and WAT. Figure 9.5 shows the schematic of this experiment; briefly tissue mitochondria were isolated and quantified against mitochondria isolated from SILAC labeled cognate cell types (3T3-L1 and brown adipocytes). The mitochondrial proteome identified and quantified in this experiment was filtered against the candidate mitochondrial catalogue established in earlier steps to arrive at high confidence *in vivo* quantitative BAT vs. WAT mitochondrial proteome of 978 proteins. The relative ratios of proteins in two tissue types provided a quantitative landscape of the differences in BAT versus WAT mitochondrial proteins, which were further shown to be involved in many biological processes and pathways. One of the key findings therein was that there is considerable specialization and divergence of key mitochondrial pathways in the two tissue types as shown in Figure 9.6. For instance, strongly up-regulated pathways in BAT mitochondria were ubiquinone biosynthesis (p=2.8E-12), oxidative phosphorylation (p=7.6E-28), and citrate cycle (p=8.6E-10). In WAT mitochondria up-regulation of pathways was less pronounced, and significantly up-regulated pathways were androgen/estrogen metabolism (p=3.8E-3), fatty acid synthesis (p=2.4E-3), pyruvate metabolism (p=1.6E-4), and metabolism of xenobiotics (p=8.4E-2).

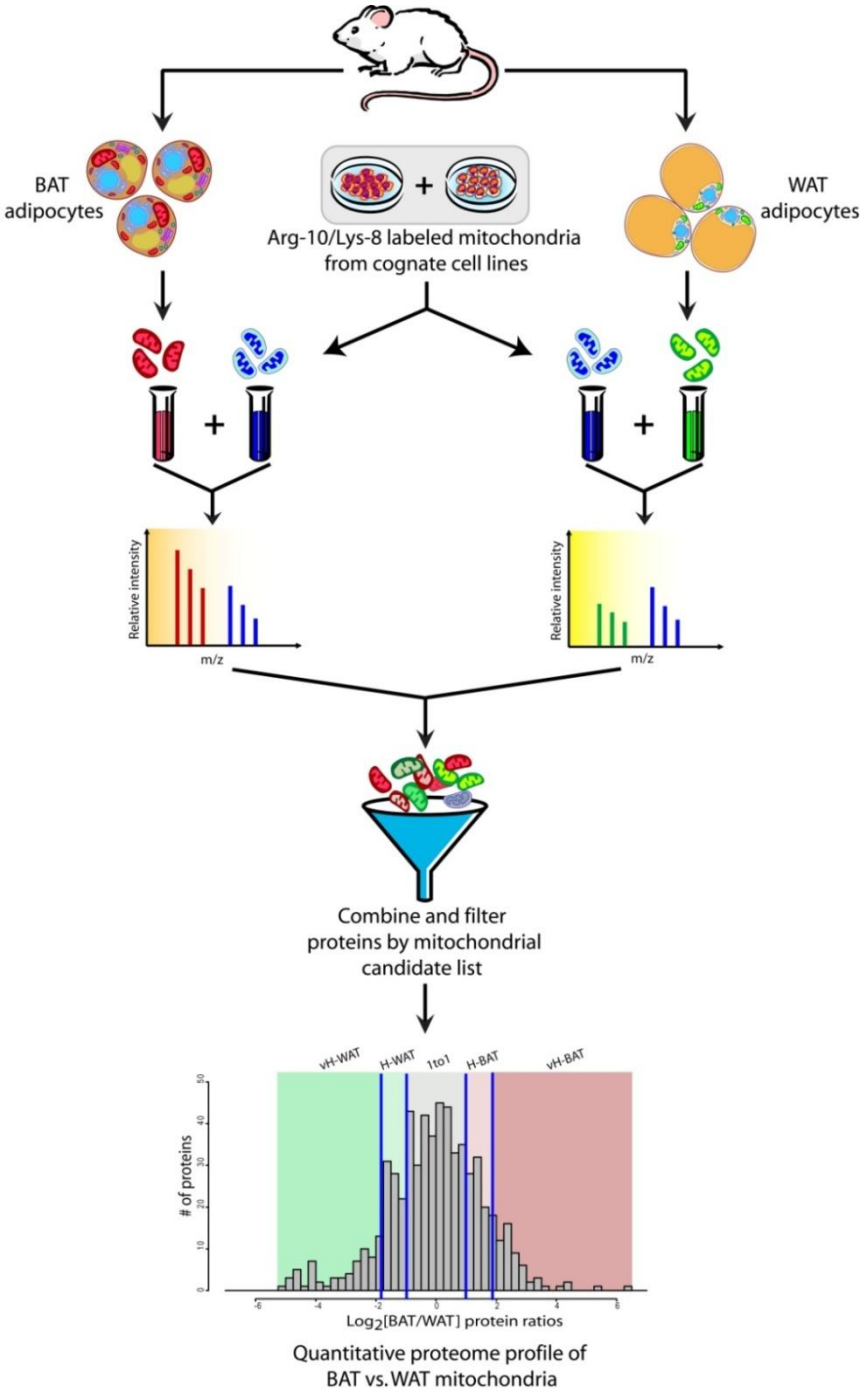


Figure 9.5 Strategy of quantitative proteomic analysis of mitochondria in BAT vs. WAT. Mitochondria isolated from brown and white adipocytes were mixed with SILAC labeled mitochondria isolated from bat and 3T3-L1 cells that served as internal standards. In vivo relative protein levels of BAT versus WAT were obtained from the "ratio of ratios" of peptide levels measured by mass spectrometry. The final dataset was the WB-mitochondrial core proteome. Proteins were qualified into one of five protein categories obtained by subdividing the BAT/WAT distribution in percentiles (10%, 25%, 75%, and 90%). The categories were named vH-BAT, H-BAT, 1to1, H-WAT, and vH-WAT to express their relative abundance in BAT versus WAT.

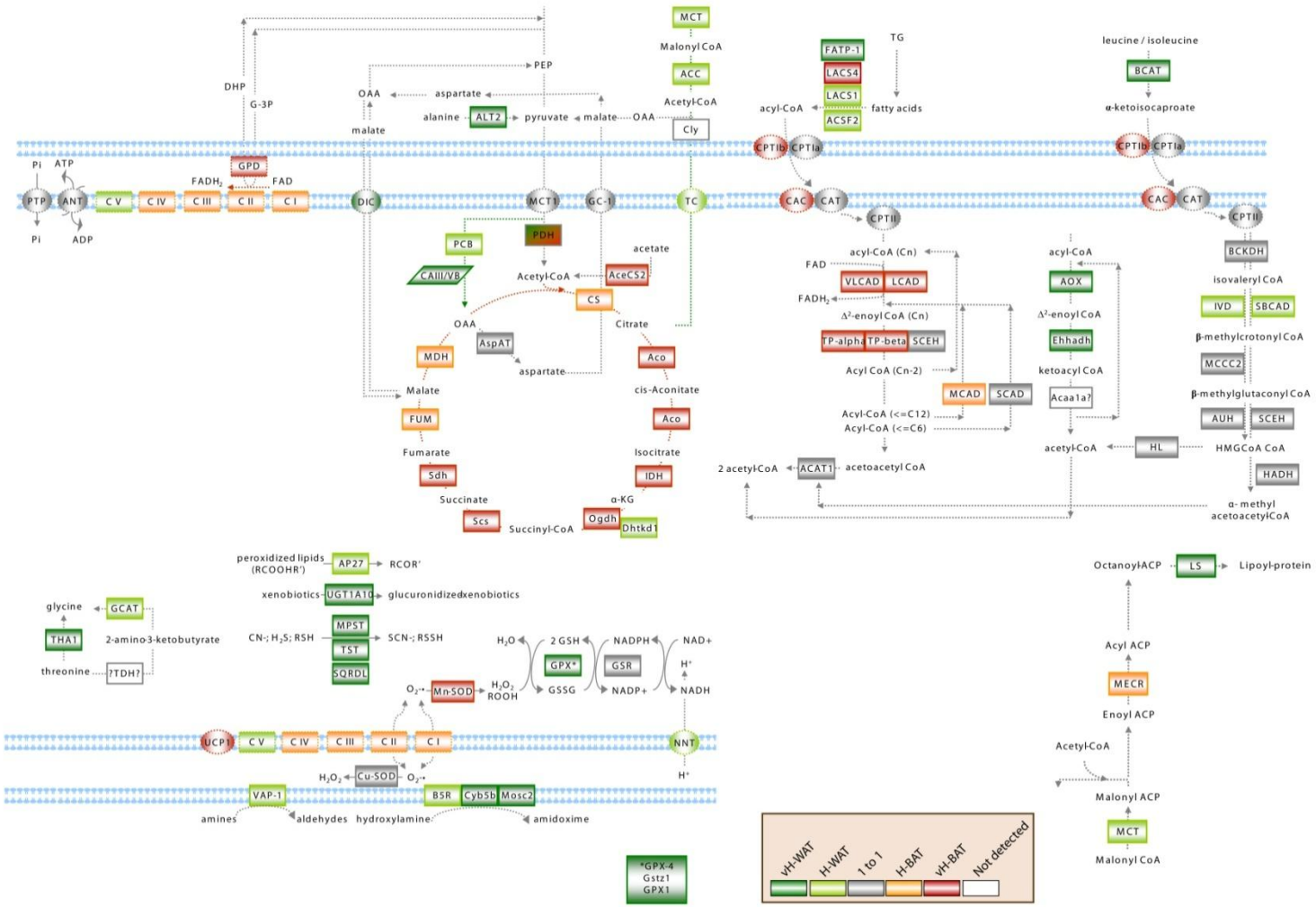


Figure 9.6 Map of metabolic pathways in primary adipocyte mitochondria from the systematic analysis of quantitative proteomic data. Relative activity of pathways in BAT and WAT mitochondria were inferred from the protein ratios obtained from quantitative proteomics(Figure 9.5). Proteins were subdivided into five categories (see Figure 2) and color-coded based on their BAT/WAT protein ratios: vH-BAT (red), H-BAT (orange), 1to1 (grey), H-WAT (green), vH-WAT (dark green). Non-filled boxes represent proteins that were not identified in our proteomic survey or that were non-mitochondrial based on our localization assignment. A tentative localization of MCC-32 with its putative interaction partners as obtained from our preliminary experiments is also shown.

In summary our framework adds to the growing toolkit of organellar proteomics and is very generic in terms of applicability. The simplicity and flexibility of our approach makes it amenable for global-study of other organelle proteomes and opens newer vistas towards creation of organelle cellular maps[334]. The framework and the mitochondrial database generated will be of great value to the ongoing systems biology and 'omics' endeavors.

10. Conclusions, challenges and perspective

Mass spectrometry based proteomics is now a multidisciplinary scientific endeavor with extensive applications and far-reaching impact. This transition from a previous niche domain to one of the strongest stakeholders of post-genomics science has been propelled by technological advances in mass spectrometry instrumentation complemented by experimental and bioinformatics innovations. A simple comparison of PubMed indexed publications appearing in last 15 years (starting from 1994 when the term "proteomics" was coined) in the field of proteomics and genomics reveals that proteomics is now as widely entrenched in contemporary biomedical research as genomics (Figure 10.1). This in itself is a testimony to the power and importance of this relatively young discipline in current scientific ecosystem.

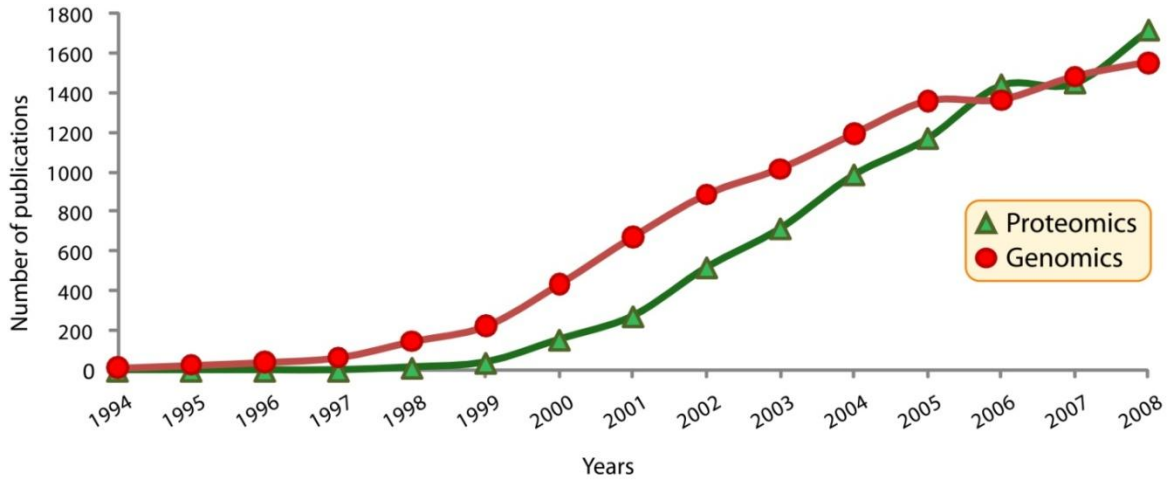


Figure 10.1 Number of publications in PubMed with title or abstract containing term "Genomics" or "Proteomics" from 1, January 1994 till 30, September 2008. The graph shows the extrapolated values for the end of the year 2008 based on the number of publications till 30, September 2008. The trend illustrates the pervasiveness of proteomics in post genomics biomedical research.

Analogous to every emerging paradigm, proteomics too has brought in its unique set of challenges that are of varying constitution - scientific, technological, experimental and computational. The very nature, scale and novelty of these challenges have attracted serious attention from the stakeholders of proteomics community. Concerted scientific and technical efforts are now underway to harness the untapped potential of proteomics.

Modern high resolution mass spectrometry instruments can produce gigabytes of data per run and a large proteomics project may employ several stages of upfront protein or peptide fractionation and consist of hundreds of runs. As proteomics endeavors become more ambitious and more comprehensive, the analytical challenges are further compounded. The high dimensionality and complexity of MS data pose novel computational, analytical and infrastructural challenges hitherto unseen by biomedical informatics researchers.

Proteomics in its current form requires extensive informatics support, therefore computational proteomics and bioinformatics have become key constituents of this field. While computational proteomics helps in extracting protein identity and quantity information from mass spectra, bioinformatics subsequently serves in discovering knowledge models, verifying hypothesis and providing biological insights. Throughout my PhD studies I have concentrated on this latter aspect of proteomics dataset analysis by bioinformatics. As proteomics progresses towards achieving the same kind of depth and comprehensiveness as genomics[2,64] it cannot be overemphasized that the analysis of current proteomics datasets necessitates elaborate bioinformatics infrastructure and support. Proteomics has opened up newer vistas for bioinformatics researcher and now drives a major part of current biomedical informatics research initiatives. In that context I have adopted two approaches towards proteomics data analysis: (1) adapting functional genomics databases, tools and algorithms for obtaining insights into proteomics dataset and, (2) developing novel analysis algorithms, frameworks and workflows for proteomics dataset. Taking specific examples of typical datasets that are being currently generated in our laboratory (beginning with qualitative catalogues to quantitative multi-time course datasets), I have tried to showcase the diversity of analysis which is needed in typical proteomics experimental scenarios. In this thesis I have discussed some of the novel analytical workflows and algorithms we have developed in our group for functional analysis of high throughput proteomics data using bioinformatics algorithms, tools and databases. All of the projects discussed in this work also showcase the importance of collaborative and interdisciplinary science wherein active dialogue and synergy is required between bioinformatics and experimental biological scientists.

Current trends in proteomics data analysis also indicate that the present bioinformatics resources and approaches will not be sufficient or adequate to mine proteomics datasets, and novel algorithms and approaches will have to be developed to integrate these datasets with disparate "omics" datasets for knowledge discovery. In that direction novel data mining, analytical and visualization software needs to be developed to harness the uniqueness of such dataset. At the same time one of the biggest challenges faced by current proteomics researcher is the relative scarcity of protein centric annotational knowledgebases. Still today most of the annotational databases are "gene" centric, and while they have been of immense value to researchers they still do not meet the numerous and at times unique demands of the proteomics community. Therefore more scientific investments are required to have a unified and comprehensive proteomics database on the lines of GenBank or Ensemble.

Modern proteomics technologies and its applications span a broad spectrum of biological explorations on various levels of cellular organization. These investigations cover nearly all aspects of cellular composition and architecture including, elucidation of structural, spatial, temporal and relational constitution - at the proteome level. The availability of such data types has in turn infused vigor into the ongoing bioinformatics efforts towards assimilating this important piece of information into the broader framework of systems biology[59,362,363]. Future bioinformatics activities in proteomics will focus more and more on integrative systems biology, as there are still many open ended questions which can only be answered by adopting this approach. For instance, we still do not have a comprehensive understanding of how protein expression is controlled and regulated as a function of regulatory mechanisms at epigenetic, transcriptional, translational and post-translational levels[215]. Current debates in biomedical informatics research are replete with many such questions.

In the post-genomic era proteomics along with other "omics" disciplines provides the foundations on which future promises of systems biology will be realized and delivered. The next steps in this direction is consolidation and integration of datasets and information across different layers of the "omics" hierarchy[364](Figure 10.2), ultimately leading to physiologically exact and clinically relevant *in silico* models of biological processes and systems. Proteomics has

the potential of making a huge impact in this endeavor by providing comprehensive and quantitative data of constituent proteomes for the systems of interest.

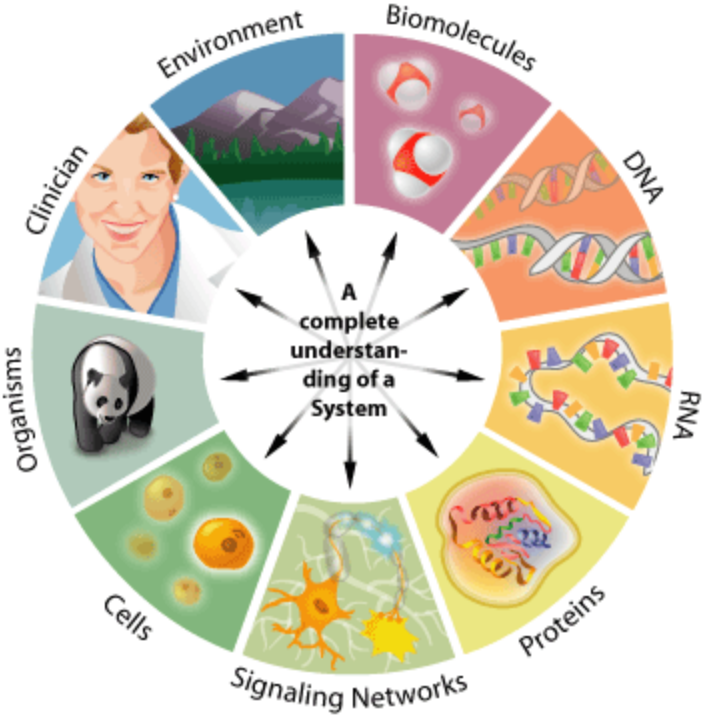


Figure 10.2. The Wheel of Biological Understanding. System biology strives to understand all aspects of an organism and its environment through the combination of a variety of scientific fields (image adapted from Joanne Fox article URL: http://bioinformatics.ubc.ca/about/what_is_bioinformatics/)

One of the ultimate litmus tests for proteomics is to be able to generate data of the nature and scale which is necessary for incorporation into multi-scale simulation and modeling frameworks[365]. Moreover, augmentation of modeling languages is needed to incorporate proteomics datasets and results into the framework of executable cell biology[366]. In that context innovative experimental strategies and scalable instrumentation capabilities have to be developed so that fine grained, comprehensive and quantitative datasets are generated in the future, which are amenable for *in silico* modeling and execution. Recent results indicate that proteomics is well equipped to handle this challenge and is poised to transform the landscape of system biology, thereby engendering profound changes in translational research.

11. Bibliography

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).
2. Cox, J. & Mann, M. Is proteomics the new genomics? *Cell* **130**, 395-398 (2007).
3. Matthiesen, R. Methods, algorithms and tools in computational proteomics: a practical point of view. *Proteomics* **7**, 2815-2832 (2007).
4. Ouzounis, C.A. & Valencia, A. Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics* **19**, 2176-2190 (2003).
5. Hagen, J.B. The origins of bioinformatics. *Nat Rev Genet* **1**, 231-236 (2000).
6. Rosen, E.D. & Spiegelman, B.M. Adipocytes as regulators of energy balance and glucose homeostasis. *Nature* **444**, 847-853 (2006).
7. Gesta, S., Tseng, Y.H. & Kahn, C.R. Developmental origin of fat: tracking obesity to its source. *Cell* **131**, 242-256 (2007).
8. Aerts, S. et al. Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**, 537-544 (2006).
9. Ong, S.E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386 (2002).
10. Warburg, O., Posener, K. & Negelein, E. Ueber den Stoffwechsel der Tumoren. *Biochemische Zeitschrift* **152**, 319-344 (1930).
11. Nobel, D. The Music of Life - Biology beyond the Genome. (Oxford University Press, 2007).
12. Watanabe, H. et al. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429**, 382-388 (2004).
13. Khaitovich, P., Paabo, S. & Weiss, G. Toward a neutral evolutionary model of gene expression. *Genetics* **170**, 929-939 (2005).
14. Clamp, M. et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* **104**, 19428-19433 (2007).
15. Xing, Y. & Lee, C. Relating alternative splicing to proteome complexity and genome evolution. *Adv Exp Med Biol* **623**, 36-49 (2007).
16. Pearson, H. Biologists initiate plan to map human proteome. *Nature* **452**, 920-921 (2008).
17. Service, R.F. Proteomics. Proteomics ponders prime time. *Science* **321**, 1758-1761 (2008).
18. Fields, S. Proteomics. Proteomics in genomeland. *Science* **291**, 1221-1224 (2001).
19. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837-846 (2000).
20. Wilkins, M.R. et al. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev* **13**, 19-50 (1996).
21. Wilkins, M.R., Williams, K.L., Appel, R.D. & Hochstrasser, D.F. (eds.) Proteome Research: New Frontiers in Functional Genomics, Edn. 1. (Springer, 1997).
22. MacBeath, G. Protein microarrays and proteomics. *Nat Genet* **32 Suppl**, 526-532 (2002).
23. Causier, B. Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrom Rev* **23**, 350-367 (2004).
24. Stevens, R.C., Yokoyama, S. & Wilson, I.A. Global efforts in structural genomics. *Science* **294**, 89-92 (2001).

25. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312**, 212-217 (2006).
26. Kelleher, N. et al. Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry. *J. Am. Chem. Soc.* **121**, 806 (1999).
27. Macek, B., Waanders, L., Olsen, J.V. & Mann, M. Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Mol Cell Proteomics* (2006).
28. Han, X., Jin, M., Breuker, K. & McLafferty, F.W. Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* **314**, 109-112 (2006).
29. de Godoy, L.M. et al. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol* **7**, R50 (2006).
30. Graumann, J. et al. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol Cell Proteomics* **7**, 672-683 (2008).
31. Liu, H., Lin, D. & Yates, J.R., 3rd Multidimensional separations for protein/peptide analysis in the post-genomic era. *BioTechniques* **32**, 898, 900, 902 passim (2002).
32. Motoyama, A., Xu, T., Ruse, C.I., Wohlschlegel, J.A. & Yates, J.R., 3rd Anion and cation mixed-bed ion exchange for enhanced multidimensional separations of peptides and phosphopeptides. *Anal Chem* **79**, 3623-3634 (2007).
33. Lu, A., Wisniewski, J.R. & Mann, M. Comparative Proteomic Profiling of Membrane Proteins in Rat Cerebellum, Spinal Cord, and Sciatic Nerve. *Manuscript to be submitted* (2008).
34. Blagoev, B. et al. A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nature biotechnology* **21**, 315-318 (2003).
35. Blagoev, B., Ong, S.E., Kratchmarova, I. & Mann, M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature biotechnology* **22**, 1139-1145 (2004).
36. Ficarro, S.B. et al. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nature biotechnology* **20**, 301-305 (2002).
37. Gruhler, A. et al. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* **4**, 310-327 (2005).
38. Larsen, M.R., Thingholm, T.E., Jensen, O.N., Roepstorff, P. & Jorgensen, T.J. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol Cell Proteomics* **4**, 873-886 (2005).
39. Thingholm, T.E., Jensen, O.N., Robinson, P.J. & Larsen, M.R. SIMAC (sequential elution from IMAC), a phosphoproteomics strategy for the rapid separation of monophosphorylated from multiply phosphorylated peptides. *Mol Cell Proteomics* **7**, 661-671 (2008).
40. Lu, A. et al. Nanoelectrospray peptide mapping revisited: Composite survey spectra allow high dynamic range protein characterization without LCMS on an orbitrap mass spectrometer. *International Journal of Mass Spectrometry* **268**, 158-167 (2007).
41. Colinge, J. & Bennett, K.L. Introduction to computational proteomics. *PLoS Comput Biol* **3**, e114 (2007).
42. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699-711 (2004).

43. Webb-Robertson, B.J. & Cannon, W.R. Current trends in computational inference from mass spectrometry-based proteomics. *Brief Bioinform* **8**, 304-317 (2007).
44. Eidhammer, I., Flikka, K., Martens, L. & Mikalsen, S. Computational Methods for Mass Spectrometry Proteomics. (Wiley, 2008).
45. Ong, S.E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1**, 252-262 (2005).
46. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods* **4**, 787-797 (2007).
47. Cox, J. & Mann, M. High peptide identification rates and proteome-wide quantitation via novel computational strategies. *Manuscript submitted* (2008).
48. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).
49. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-280 (2004).
50. Adachi, J., Kumar, C., Zhang, Y. & Mann, M. In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol Cell Proteomics* **6**, 1257-1273 (2007).
51. Adachi, J., Kumar, C., Zhang, Y., Olsen, J.V. & Mann, M. The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol* **7**, R80 (2006).
52. Brunner, E. et al. A high-quality catalog of the Drosophila melanogaster proteome. *Nature biotechnology* **25**, 576-583 (2007).
53. Forner, F., Kumar, C., Luber, C.A. & Mann, M. Pathway analysis of mitochondria in brown versus white adipocytes by quantitative proteomics. *Manuscript submitted* (2008).
54. Pan, C., Kumar, C., Bohl, S., Klingmuller, U. & Mann, M. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type specific functions. *Manuscript submitted* (2008).
55. Olsen, J.V. et al. A systems view of the cell cycle by quantitative phosphoproteomics. *Manuscript submitted* (2008).
56. Ideker, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-934 (2001).
57. Mootha, V.K. et al. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629-640 (2003).
58. Foster, L.J. et al. A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187-199 (2006).
59. Pagliarini, D.J. et al. A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134**, 112-123 (2008).
60. Prokisch, H. et al. Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol* **2**, e160 (2004).
61. Hood, L., Heath, J.R., Phelps, M.E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640-643 (2004).
62. Joyce, A.R. & Palsson, B.O. The model organism as a system: integrating 'omics' data sets. *Nature reviews* **7**, 198-210 (2006).
63. Smith, J.C. & Figeys, D. Proteomics technology in systems biology. *Mol Biosyst* **2**, 364-370 (2006).

64. de Godoy, L.M. et al. Comprehensive, mass spectrometry-based proteome quantitation of haploid versus diploid yeast. *Nature(accepted)* (2008).
65. Olsen, J.V. et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648 (2006).
66. Clegg, G.A. & Dole, M. Molecular beams of macroions. 3. Zein and polyvinylpyrrolidone. *Biopolymers* **10**, 821-826.
67. Whitehouse, C.M., Dreyer, R.N., Yamashita, M. & Fenn, J.B. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem* **57**, 675-679 (1985).
68. Kebarle, P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *J Mass Spectrom* **35**, 804-817 (2000).
69. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-2301 (1988).
70. Tanaka, K. The origin of macromolecule ionization by laser irradiation (Nobel lecture). *Angew Chem Int Ed Engl* **42**, 3860-3870 (2003).
71. van Berkel, W.J., van den Heuvel, R.H., Versluis, C. & Heck, A.J. in Protein Sci, Vol. 9 435-4392000).
72. Cotter, R.J. Time-of-flight mass spectrometry: an increasing role in the life sciences. *Biomed Environ Mass Spectrom* **18**, 513-532 (1989).
73. Hager, J.W. & Le Blanc, J.C. High-performance liquid chromatography-tandem mass spectrometry with a new quadrupole/linear ion trap instrument. *J Chromatogr A* **1020**, 3-9 (2003).
74. Schwartz, J.C., Senko, M.W. & Syka, J.E. A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* **13**, 659-669 (2002).
75. Morris, H.R. et al. High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* **10**, 889-896 (1996).
76. Shevchenko, A. et al. Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* **11**, 1015-1024 (1997).
77. Guan, S., Marshall, A.G. & Wahl, M.C. MS/MS with high detection efficiency and mass resolving power for product ions in Fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* **66**, 1363-1367 (1994).
78. Martin, S.E., Shabanowitz, J., Hunt, D.F. & Marto, J.A. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* **72**, 4266-4274 (2000).
79. Hunt, D.F. et al. Tandem quadrupole-Fourier transform mass spectrometry of oligopeptides. *Anal Chem* **57**, 2728-2733 (1985).
80. Syka, J.E. et al. Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res* **3**, 621-626 (2004).
81. Glish, G.L. & Vachet, R.W. The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov* **2**, 140-150 (2003).
82. Hinsby, A.M., Olsen, J.V. & Mann, M. Tyrosine phosphoproteomics of fibroblast growth factor signaling: a role for insulin receptor substrate-4. *J Biol Chem* **279**, 46438-46447 (2004).

83. Dieguez-Acuna, F.J. et al. Characterization of mouse spleen cells by subtractive proteomics. *Mol Cell Proteomics* **4**, 1459-1470 (2005).
84. Olsen, J.V., Ong, S.E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics* **3**, 608-614 (2004).
85. Hu, Q. et al. The Orbitrap: a new mass spectrometer. *J Mass Spectrom* **40**, 430-443 (2005).
86. Olsen, J.V. et al. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* **4**, 2010-2021 (2005).
87. Mann, M. Quantitative proteomics? *Nat Biotechnol* **17**, 954-955 (1999).
88. Jonckheere, J.A. et al. Selected ion monitoring assay for bromhexine in biological fluids. *Biomed Mass Spectrom* **7**, 582-587 (1980).
89. Gygi, S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology* **17**, 994-999 (1999).
90. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96**, 6591-6596 (1999).
91. Yao, X., Freas, A., Ramirez, J., Demirev, P.A. & Fenselau, C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-2842 (2001).
92. Ross, P.L. et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-1169 (2004).
93. Olsen, J.V. et al. HysTag--a novel proteomic quantification tool applied to differential display analysis of membrane proteins from distinct areas of mouse brain. *Mol Cell Proteomics* **3**, 82-92 (2004).
94. Lahm, H.W. & Langen, H. Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis* **21**, 2105-2114 (2000).
95. Krijgsveld, J. et al. Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nature biotechnology* **21**, 927-931 (2003).
96. Wu, C.C., MacCoss, M.J., Howell, K.E., Matthews, D.E. & Yates, J.R., 3rd Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal Chem* **76**, 4951-4959 (2004).
97. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
98. Kruger, M. et al. SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* **134**, 353-364 (2008).
99. Stewart, II, Thomson, T. & Figeys, D. 18O labeling: a tool for proteomics. *Rapid Commun Mass Spectrom* **15**, 2456-2465 (2001).
100. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **100**, 6940-6945 (2003).
101. Beynon, R.J., Doherty, M.K., Pratt, J.M. & Gaskell, S.J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat Methods* **2**, 587-589 (2005).
102. Hanke, S., Besir, H., Oesterhelt, D. & Mann, M. Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *J Proteome Res* **7**, 1118-1130 (2008).

103. Durr, E. et al. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nature biotechnology* **22**, 985-992 (2004).
104. Andersen, J.S. et al. Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**, 570-574 (2003).
105. Domon, B. & Aebersold, R. Challenges and opportunities in proteomics data analysis. *Mol Cell Proteomics* **5**, 1921-1926 (2006).
106. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999).
107. Eng, J., McCormack, A.L. & Yates, J.R., III. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **80**, 976-989 (1994).
108. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P.A. Protein identification by spectral networks analysis. *Proc Natl Acad Sci U S A* **104**, 6140-6145 (2007).
109. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658 (2003).
110. Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P. & Gygi, S.P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature biotechnology* **22**, 214-219 (2004).
111. Jensen, O.N. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol* **7**, 391-403 (2006).
112. Mann, M. & Jensen, O.N. Proteomic analysis of post-translational modifications. *Nature biotechnology* **21**, 255-261 (2003).
113. Savitski, M.M., Nielsen, M.L. & Zubarev, R.A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics* **5**, 935-948 (2006).
114. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **389**, 1017-1031 (2007).
115. Malmstrom, J., Lee, H. & Aebersold, R. Advances in proteomic workflows for systems biology. *Curr Opin Biotechnol* **18**, 378-384 (2007).
116. Patterson, S.D. & Aebersold, R.H. Proteomics: the first decade and beyond. *Nat Genet* **33 Suppl**, 311-323 (2003).
117. Pang, J.X., Ginanni, N., Dongre, A.R., Hefta, S.A. & Opitek, G.J. Biomarker discovery in urine by proteomics. *J Proteome Res* **1**, 161-169 (2002).
118. Gao, J., Opiteck, G.J., Friedrichs, M.S., Dongre, A.R. & Hefta, S.A. Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res* **2**, 643-649 (2003).
119. Liu, H., Sadygov, R.G. & Yates, J.R., 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193-4201 (2004).
120. Listgarten, J. & Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* **4**, 419-434 (2005).
121. Taylor, C.F. et al. The minimum information about a proteomics experiment (MIAPE). *Nature biotechnology* **25**, 887-893 (2007).

122. Astbury, W.T. Adventures in molecular biology. *Harvey Lect* **Series 46**, 3-44 (1950).
123. Astbury, W.T. Molecular biology or ultrastructural biology? *Nature* **190**, 1124 (1961).
124. Sanger, F. The arrangement of amino acids in proteins. *Adv Protein Chem* **7**, 1-67 (1952).
125. Laskowski, R.A. & Thornton, J.M. Understanding the molecular machinery of genetics through 3D structures. *Nat Rev Genet* **9**, 141-151 (2008).
126. Kay, L.E. Who wrote the book of life? Information and the transformation of molecular biology, 1945-55. *Science in Context* **8**, 609-634 (1995).
127. Kay, L.E. Cybernetics, information, life: The emergence of scriptural representations of heredity. *Configurations* **5**, 23-91 (1997).
128. Watson, J.D. & Crick, F.H. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964-967 (1953).
129. Gamow, G., Rich, A. & Ycas, M. The problem of information transfer from the nucleic acids to proteins. *Adv Biol Med Phys* **4**, 23-68 (1956).
130. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).
131. Pauling, L., Corey, R.B. & Branson, H.R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37**, 205-211 (1951).
132. Anfinsen, C.B. & Scheraga, H.A. Experimental and theoretical aspects of protein folding. *Adv Protein Chem* **29**, 205-300 (1975).
133. Pauling, L. & Corey, R.B. Two Pleated-Sheet Configurations of Polypeptide Chains Involving Both Cis and Trans Amide Groups. *Proc Natl Acad Sci U S A* **39**, 247-252 (1953).
134. Szent-Gyorgyi, A.G. & Cohen, C. Role of proline in polypeptide chain configuration of proteins. *Science* **126**, 697-698 (1957).
135. Crick, F.H.C. THE PACKING OF ALPHA-HELICES - SIMPLE COILED-COILS. *Acta Crystallographica* **6**, 689-697 (1953).
136. Horowitz, N.H. On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci U S A* **31**, 153-157 (1945).
137. Britten, R.J. & Davidson, E.H. Gene regulation for higher cells: a theory. *Science* **165**, 349-357 (1969).
138. Turing, A.M. THE CHEMICAL BASIS OF MORPHOGENESIS. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **237**, 37-72 (1952).
139. Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546, 563-547 (1996).
140. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
141. Luscombe, N.M., Greenbaum, D. & Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* **40**, 346-358 (2001).
142. Baldi, P. & Brunak, S. Bioinformatics - The Machine Learning Approach, Edn. 1. (The MIT Press, 2001).
143. Bourne, P.E. & Weissig, H. (eds.) Structural Bioinformatics. (Wiley, 2003).
144. Miller, W., Makova, K.D., Nekrutenko, A. & Hardison, R.C. Comparative genomics. *Annu Rev Genomics Hum Genet* **5**, 15-56 (2004).
145. Vukmirovic, O.G. & Tilghman, S.M. Exploring genome space. *Nature* **405**, 820-822 (2000).

146. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113 (2004).
147. Kitano, H. Systems biology: a brief overview. *Science (New York, N.Y* **295**, 1662-1664 (2002).
148. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **8**, 995-1005 (2007).
149. Lengauer, T. (ed.) Bioinformatics - From Genomes to Therapies Vol. 1: The Building Blocks: Molecular Sequences and Structures. (Wiley, 2007).
150. Lengauer, T. (ed.) Bioinformatics - From Genomes to Therapies Vol. 2: Getting at the Inner Workings: Molecular Interactions. (Wiley, 2007).
151. Lengauer, T. (ed.) Bioinformatics - From Genomes to Therapies Vol. 3: The Holy Grail: Molecular Function. (Wiley, 2007).
152. Dunn, M.J., Jorde, L.B., Little, P.F.R. & Subramaniam, S. (eds.) Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. (Wiley, 2005).
153. Stoughton, R.B. Applications of DNA microarrays in biology. *Annu Rev Biochem* **74**, 53-82 (2005).
154. Hoheisel, J.D. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7**, 200-210 (2006).
155. Trevino, V., Falciani, F. & Barrera-Saldana, H.A. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med* **13**, 527-541 (2007).
156. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126 (2000).
157. Clarke, R. et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* **8**, 37-49 (2008).
158. Golub, T.R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).
159. Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. & Chinnaiyan, A.M. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* **62**, 4427-4433 (2002).
160. Alizadeh, A.A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).
161. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).
162. Tamayo, P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* **96**, 2907-2912 (1999).
163. Lapointe, J. et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* **101**, 811-816 (2004).
164. van 't Veer, L.J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).
165. Sorlie, T., Sexton, H.C., Busund, R. & Sorlie, D. A global measure of physical functioning: psychometric properties. *Health Serv Res* **36**, 1109-1124 (2001).
166. Ramaswamy, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98**, 15149-15154 (2001).
167. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601-620 (2000).

168. Liang, S., Fuhrman, S. & Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, 18-29 (1998).
169. Butte, A.J. & Kohane, I.S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 418-429 (2000).
170. Tringe, S.G., Wagner, A. & Ruby, S.W. Enriching for direct regulatory targets in perturbed gene-expression profiles. *Genome Biol* **5**, R29 (2004).
171. D'Haeseleer, P., Liang, S. & Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16**, 707-726 (2000).
172. Csete, M.E. & Doyle, J.C. Reverse engineering of biological complexity. *Science* **295**, 1664-1669 (2002).
173. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799-805 (2004).
174. Mischel, P.S., Cloughesy, T.F. & Nelson, S.F. DNA-microarray analysis of brain cancer: molecular classification for therapy. *Nat Rev Neurosci* **5**, 782-792 (2004).
175. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17 Suppl 1**, S215-224 (2001).
176. Maston, G.A., Evans, S.K. & Green, M.R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29-59 (2006).
177. Davidson, E. The Regulatory Genome: Gene Regulatory Networks In Development And Evolution. (Academic Press, 2006).
178. Olson, E.N. Gene regulatory networks in the evolution and development of the heart. *Science* **313**, 1922-1927 (2006).
179. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276-287 (2004).
180. Rao, C.V. & Arkin, A.P. Control motifs for intracellular regulatory networks. *Annu Rev Biomed Eng* **3**, 391-419 (2001).
181. Bolouri, H. & Davidson, E.H. Modeling transcriptional regulatory networks. *Bioessays* **24**, 1118-1129 (2002).
182. Kaern, M., Blake, W.J. & Collins, J.J. The engineering of gene regulatory networks. *Annu Rev Biomed Eng* **5**, 179-206 (2003).
183. Sheng, Y., Engstrom, P.G. & Lenhard, B. Mammalian microRNA prediction through a support vector machine model of sequence and structure. *PLoS ONE* **2**, e946 (2007).
184. Miranda, K.C. et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203-1217 (2006).
185. Bentwich, I. Identifying human microRNAs. *Curr Top Microbiol Immunol* **320**, 257-269 (2008).
186. Bock, C. & Lengauer, T. Computational epigenetics. *Bioinformatics* **24**, 1-10 (2008).
187. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686 (1997).
188. Carpenter, A.E. & Sabatini, D.M. Systematic genome-wide screens of gene function. *Nat Rev Genet* **5**, 11-22 (2004).
189. de Hoog, C.L. & Mann, M. Proteomics. *Annu Rev Genomics Hum Genet* **5**, 267-293 (2004).
190. Xia, Y. et al. Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem* **73**, 1051-1087 (2004).

191. Ma'ayan, A., Blitzer, R.D. & Iyengar, R. Toward predictive models of mammalian cells. *Annu Rev Biophys Biomol Struct* **34**, 319-349 (2005).
192. Alon, U. Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**, 450-461 (2007).
193. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555 (2002).
194. Kitano, H. Biological robustness. *Nat Rev Genet* **5**, 826-837 (2004).
195. Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* **298**, 824-827 (2002).
196. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41-42 (2001).
197. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).
198. Lee, T.I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).
199. Luscombe, N.M. et al. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308-312 (2004).
200. de Lichtenberg, U., Jensen, L.J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724-727 (2005).
201. Goh, K.I. et al. The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-8690 (2007).
202. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L. & Vidal, M. Drug-target network. *Nature biotechnology* **25**, 1119-1126 (2007).
203. Lee, D.S. et al. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A* **105**, 9880-9885 (2008).
204. Price, N.D., Reed, J.L. & Palsson, B.O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2**, 886-897 (2004).
205. Jamshidi, N. & Palsson, B.O. Formulating genome-scale kinetic models in the post-genome era. *Mol Syst Biol* **4**, 171 (2008).
206. Duarte, N.C. et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* **104**, 1777-1782 (2007).
207. Bonneau, R. et al. A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131**, 1354-1365 (2007).
208. Lee, I. et al. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet* **40**, 181-188 (2008).
209. Becker, S.A. et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* **2**, 727-738 (2007).
210. Shi, R. et al. Analysis of the mouse liver proteome using advanced mass spectrometry. *J Proteome Res* **6**, 2963-2972 (2007).
211. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **30**, 276-280 (2002).
212. Mulder, N.J. et al. New developments in the InterPro database. *Nucleic Acids Res* **35**, D224-228 (2007).
213. Pasini, E.M. et al. In-depth analysis of the membrane and cytosolic proteome of red blood cells. *Blood* **108**, 791-801 (2006).

214. Wu, L. et al. Global survey of human T leukemic cells by integrating proteomics and transcriptomics profiling. *Mol Cell Proteomics* **6**, 1343-1353 (2007).
215. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology* **25**, 117-124 (2007).
216. Lage, K. et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology* **25**, 309-316 (2007).
217. Desiere, F. et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol* **6**, R9 (2005).
218. Gupta, N. et al. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res* **17**, 1362-1377 (2007).
219. Baerenfaller, K. et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **320**, 938-941 (2008).
220. Meunier, B. et al. Assessment of hierarchical clustering methodologies for proteomic data mining. *J Proteome Res* **6**, 358-366 (2007).
221. Dunkley, T.P. et al. Mapping the Arabidopsis organelle proteome. *Proc Natl Acad Sci U S A* **103**, 6518-6523 (2006).
222. Rinner, O. et al. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nature biotechnology* **25**, 345-352 (2007).
223. Petricoin, E.F. et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572-577 (2002).
224. Ray, S. et al. Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. *Nat Med* **13**, 1359-1362 (2007).
225. Biddinger, S.B. & Kahn, C.R. From mice to men: insights into the insulin resistance syndromes. *Annu Rev Physiol* **68**, 123-158 (2006).
226. Kopelman, P.G. Obesity as a medical problem. *Nature* **404**, 635-643 (2000).
227. Rajala, M.W. & Scherer, P.E. Minireview: The adipocyte--at the crossroads of energy homeostasis, inflammation, and atherosclerosis. *Endocrinology* **144**, 3765-3773 (2003).
228. Murphy, D.J. & Vance, J. Mechanisms of lipid-body formation. *Trends Biochem Sci* **24**, 109-115 (1999).
229. Cermelli, S., Guo, Y., Gross, S.P. & Welte, M.A. The lipid-droplet proteome reveals that droplets are a protein-storage depot. *Curr Biol* **16**, 1783-1795 (2006).
230. Kratchmarova, I. et al. A proteomic approach for identification of secreted proteins during the differentiation of 3T3-L1 preadipocytes to adipocytes. *Mol Cell Proteomics* **1**, 213-222 (2002).
231. MacDougald, O.A., Cornelius, P., Lin, F.T., Chen, S.S. & Lane, M.D. Glucocorticoids reciprocally regulate expression of the CCAAT/enhancer-binding protein alpha and delta genes in 3T3-L1 adipocytes and white adipose tissue. *J Biol Chem* **269**, 19041-19047 (1994).
232. Thurmond, D.C. et al. Regulation of insulin-stimulated GLUT4 translocation by Munc18c in 3T3L1 adipocytes. *J Biol Chem* **273**, 33876-33883 (1998).
233. Piper, R.C., Hess, L.J. & James, D.E. Differential sorting of two glucose transporters expressed in insulin-sensitive cells. *Am J Physiol* **260**, C570-580 (1991).
234. Wilm, M. et al. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466-469 (1996).

235. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**, 663-670 (2003).
236. Olsen, J.V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* **101**, 13417-13422 (2004).
237. Kristensen, D.B. et al. Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol Cell Proteomics* **3**, 1023-1038 (2004).
238. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-3449 (2005).
239. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).
240. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676 (2005).
241. Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18**, 207-208 (2002).
242. Foster, L.J. et al. A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187-199 (2006).
243. Kislinger, T. et al. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125**, 173-186 (2006).
244. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. & Conklin, B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* **31**, 19-20 (2002).
245. Elias, J.E., Haas, W., Faherty, B.K. & Gygi, S.P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* **2**, 667-675 (2005).
246. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**, 1419-1440 (2005).
247. Jiang, X.S. et al. A high-throughput approach for subcellular proteome: identification of rat liver proteins using subcellular fractionation coupled with two-dimensional liquid chromatography tandem mass spectrometry and bioinformatic analysis. *Mol Cell Proteomics* **3**, 441-455 (2004).
248. Schirle, M., Heurtier, M.A. & Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* **2**, 1297-1305 (2003).
249. Soukas, A., Socci, N.D., Saatkamp, B.D., Novelli, S. & Friedman, J.M. Distinct transcriptional profiles of adipogenesis in vivo and in vitro. *J Biol Chem* **276**, 34167-34174 (2001).
250. Burton, G.R., Nagarajan, R., Peterson, C.A. & McGehee, R.E., Jr. Microarray analysis of differentiation-specific gene expression during 3T3-L1 adipogenesis. *Gene* **329**, 167-185 (2004).
251. Gerhold, D.L. et al. Gene expression profile of adipocyte differentiation and its regulation by peroxisome proliferator-activated receptor-gamma agonists. *Endocrinology* **143**, 2106-2118 (2002).

252. Hackl, H. et al. Molecular processes during fat cell development revealed by gene expression profiling and functional annotation. *Genome Biol* **6**, R108 (2005).
253. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067 (2004).
254. de Godoy, L.M. et al. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol* **7**, R50 (2005).
255. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).
256. Brooks, P. et al. Subcellular localization of proteasomes and their regulatory complexes in mammalian cells. *Biochem J* **346 Pt 1**, 155-161 (2000).
257. Ouzounis, C.A., Coulson, R.M., Enright, A.J., Kunin, V. & Pereira-Leal, J.B. Classification schemes for protein structure and function. *Nat Rev Genet* **4**, 508-519 (2003).
258. Seet, B.T., Dikic, I., Zhou, M.M. & Pawson, T. Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* **7**, 473-483 (2006).
259. Galinier, A. et al. Adipose tissue proadipogenic redox changes in obesity. *J Biol Chem* **281**, 12682-12687 (2006).
260. Urakawa, H. et al. Oxidative stress is associated with adiposity and insulin resistance in men. *J Clin Endocrinol Metab* **88**, 4673-4676 (2003).
261. Siddiqui, A.S. et al. A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc Natl Acad Sci U S A* **102**, 18485-18490 (2005).
262. Bhattacharjee, Y. Neuroscience. 'Google of the brain': atlas maps brain's genetic activity. *Science* **313**, 1879 (2006).
263. Birnbaum, M.J. Identification of a novel gene encoding an insulin-responsive glucose transporter protein. *Cell* **57**, 305-315 (1989).
264. Charron, M.J., Brosius, F.C., 3rd, Alper, S.L. & Lodish, H.F. A glucose transport protein expressed predominately in insulin-responsive tissues. *Proc Natl Acad Sci U S A* **86**, 2535-2539 (1989).
265. Fukumoto, H. et al. Cloning and characterization of the major insulin-responsive glucose transporter expressed in human skeletal muscle and other insulin-responsive tissues. *J Biol Chem* **264**, 7776-7779 (1989).
266. James, D.E., Strube, M. & Mueckler, M. Molecular cloning and characterization of an insulin-regulatable glucose transporter. *Nature* **338**, 83-87 (1989).
267. Kaestner, K.H. et al. Sequence, tissue distribution, and differential expression of mRNA for a putative insulin-responsive glucose transporter in mouse 3T3-L1 adipocytes. *Proc Natl Acad Sci U S A* **86**, 3150-3154 (1989).
268. Miinea, C.P. et al. AS160, the Akt substrate regulating GLUT4 translocation, has a functional Rab GTPase-activating protein domain. *Biochem J* **391**, 87-93 (2005).
269. Larance, M. et al. Characterization of the role of the Rab GTPase-activating protein AS160 in insulin-regulated GLUT4 trafficking. *J Biol Chem* **280**, 37803-37813 (2005).
270. Imamura, T. et al. Insulin-induced GLUT4 translocation involves protein kinase C-lambda-mediated functional coupling between Rab4 and the motor protein kinesin. *Mol Cell Biol* **23**, 4892-4900 (2003).

271. Millar, C.A., Powell, K.A., Hickson, G.R., Bader, M.F. & Gould, G.W. Evidence for a role for ADP-ribosylation factor 6 in insulin-stimulated glucose transporter-4 (GLUT4) trafficking in 3T3-L1 adipocytes. *J Biol Chem* **274**, 17619-17625 (1999).
272. Usui, I., Imamura, T., Huang, J., Satoh, H. & Olefsky, J.M. Cdc42 is a Rho GTPase family member that can mediate insulin signaling to glucose transport in 3T3-L1 adipocytes. *J Biol Chem* **278**, 13765-13774 (2003).
273. Zhang, Y. et al. MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes. *Nucleic Acids Res* **35**, D771-779 (2007).
274. Freshney, R.I. Culture of Animal Cells: A Manual of Basic Technique, Edn. 5th Edition. (Wiley, 2005).
275. Masters, J.R. HeLa cells 50 years on: the good, the bad and the ugly. *Nat Rev Cancer* **2**, 315-319 (2002).
276. Masters, J.R. Human cancer cell lines: fact and fantasy. *Nat Rev Mol Cell Biol* **1**, 233-236 (2000).
277. Burdall, S.E., Hanby, A.M., Lansdown, M.R. & Speirs, V. Breast cancer cell lines: friend or foe? *Breast Cancer Res* **5**, 89-95 (2003).
278. Nolan, G.P. What's wrong with drug screening today. *Nat Chem Biol* **3**, 187-191 (2007).
279. Kamb, A. What's wrong with our cancer models? *Nat Rev Drug Discov* **4**, 161-165 (2005).
280. Sandberg, R. & Ernberg, I. The molecular portrait of in vitro growth by meta-analysis of gene-expression profiles. *Genome Biol* **6**, R65 (2005).
281. Olsavsky, K.M. et al. Gene expression profiling and differentiation assessment in primary human hepatocyte cultures, established hepatoma cell lines, and human liver tissues. *Toxicol Appl Pharmacol* **222**, 42-56 (2007).
282. Klingmuller, U. et al. Primary mouse hepatocytes for systems biology approaches: a standardized in vitro system for modelling of signal transduction pathways. *Syst Biol (Stevenage)* **153**, 433-447 (2006).
283. Graumann, J. et al. SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins. *Mol Cell Proteomics* (2007).
284. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)* **23**, 257-258 (2007).
285. Team, R.D.C. R: a language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria; 2004).
286. Mann, M. Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol* **7**, 952-958 (2006).
287. Darlington, G.J., Bernhard, H.P., Miller, R.A. & Ruddle, F.H. Expression of liver phenotypes in cultured mouse hepatoma cells. *J Natl Cancer Inst* **64**, 809-819 (1980).
288. Ashburner, M. & Lewis, S. On ontologies for biologists: the Gene Ontology--untangling the web. *Novartis Foundation symposium* **247**, 66-80; discussion 80-63, 84-90, 244-252 (2002).
289. Aitken, A.E., Richardson, T.A. & Morgan, E.T. Regulation of drug-metabolizing enzymes and transporters in inflammation. *Annu Rev Pharmacol Toxicol* **46**, 123-149 (2006).
290. Kawajiri, K. & Fujii-Kuriyama, Y. Cytochrome P450 gene regulation and physiological functions mediated by the aryl hydrocarbon receptor. *Arch Biochem Biophys* **464**, 207-212 (2007).

291. Rivera, S.P., Saarikoski, S.T. & Hankinson, O. Identification of a novel dioxin-inducible cytochrome P450. *Mol Pharmacol* **61**, 255-259 (2002).
292. Hewitt, N.J. et al. Primary hepatocytes: current understanding of the regulation of metabolic enzymes and transporter proteins, and pharmaceutical practice for the use of hepatocytes in metabolism, enzyme induction, transporter, clearance, and hepatotoxicity studies. *Drug Metab Rev* **39**, 159-234 (2007).
293. Guidotti, J.E. et al. Liver cell polyploidization: a pivotal role for binuclear hepatocytes. *J Biol Chem* **278**, 19095-19101 (2003).
294. Jakowlew, S.B. Transforming growth factor-beta in cancer and metastasis. *Cancer Metastasis Rev* **25**, 435-457 (2006).
295. Butte, A. The use and analysis of microarray data. *Nat Rev Drug Discov* **1**, 951-960 (2002).
296. Kastan, M.B. & Bartek, J. Cell-cycle checkpoints and cancer. *Nature* **432**, 316-323 (2004).
297. Morgen, D.O. The Cell Cycle; Principles of Control. (New Science Press Ltd, 2007).
298. Whitfield, M.L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* **13**, 1977-2000 (2002).
299. Blangy, A., Arnaud, L. & Nigg, E.A. Phosphorylation by p34cdc2 protein kinase regulates binding of the kinesin-related motor HsEg5 to the dynactin subunit p150. *J Biol Chem* **272**, 19418-19424 (1997).
300. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **1**, 2856-2860 (2006).
301. Schroeder, M.J., Shabanowitz, J., Schwartz, J.C., Hunt, D.F. & Coon, J.J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem* **76**, 3590-3598 (2004).
302. Gnad, F. et al. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* **8**, R250 (2007).
303. Mardia, K.V., Jupp, P. E. Directional Statistics, Edn. Second. (John Wiley and Sons Ltd, 2000).
304. Carson, J.P. et al. Pharmacogenomic identification of targets for adjuvant therapy with the topoisomerase poison camptothecin. *Cancer Res* **64**, 2096-2104 (2004).
305. Team, R.D.C. R: a language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria; 2008).
306. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).
307. Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. & Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633-1649 (2004).
308. Linding, R. et al. NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic acids research* **36**, D695-699 (2008).
309. Kittler, R. et al. Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nature cell biology* **9**, 1401-1412 (2007).
310. Orlando, D.A. et al. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature* **453**, 944-947 (2008).

311. Black, J.D. Protein kinase C-mediated regulation of the cell cycle. *Front Biosci* **5**, D406-423 (2000).
312. Matsuoka, S. et al. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science (New York, N.Y* **316**, 1160-1166 (2007).
313. Harper, J.W. & Elledge, S.J. The DNA damage response: ten years after. *Molecular cell* **28**, 739-745 (2007).
314. Branzei, D. & Foiani, M. Regulation of DNA repair throughout the cell cycle. *Nature reviews* **9**, 297-308 (2008).
315. Cortez, D., Glick, G. & Elledge, S.J. Minichromosome maintenance proteins are direct targets of the ATM and ATR checkpoint kinases. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 10078-10083 (2004).
316. Ishimi, Y. A DNA helicase activity is associated with an MCM4, -6, and -7 protein complex. *The Journal of biological chemistry* **272**, 24508-24513 (1997).
317. Neumann, B. et al. High-throughput RNAi screening by time-lapse imaging of live human cells. *Nature methods* **3**, 385-390 (2006).
318. Furukawa, K. & Kondo, T. Identification of the lamina-associated-polypeptide-2-binding domain of B-type lamin. *European journal of biochemistry / FEBS* **251**, 729-733 (1998).
319. Foisner, R. & Gerace, L. Integral membrane proteins of the nuclear envelope interact with lamins and chromosomes, and binding is modulated by mitotic phosphorylation. *Cell* **73**, 1267-1279 (1993).
320. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).
321. Dorner, D. et al. Lamina-associated polypeptide 2alpha regulates cell cycle progression and differentiation via the retinoblastoma-E2F pathway. *The Journal of cell biology* **173**, 83-93 (2006).
322. Huh, W.K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691 (2003).
323. Seibel, N.M., Eljouni, J., Nalaskowski, M.M. & Hampe, W. Nuclear localization of enhanced green fluorescent protein homomultimers. *Analytical biochemistry* **368**, 95-99 (2007).
324. Berglund, L. et al. A gene-centric human protein atlas for expression profiles based on antibodies. *Mol Cell Proteomics* (2008).
325. Nakai, K. & Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14**, 897-911 (1992).
326. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O. & Ofran, Y. Automatic prediction of protein function. *Cell Mol Life Sci* **60**, 2637-2650 (2003).
327. Donnes, P. & Hoglund, A. Predicting protein subcellular localization: past, present, and future. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute* **2**, 209-215 (2004).
328. Gardy, J.L. & Brinkman, F.S. Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol* **4**, 741-751 (2006).
329. Sprenger, J., Fink, J.L. & Teasdale, R.D. Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC bioinformatics* **7 Suppl 5**, S3 (2006).
330. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols* **2**, 953-971 (2007).

331. Habib, S.J., Neupert, W. & Rapaport, D. Analysis and prediction of mitochondrial targeting signals. *Methods in cell biology* **80**, 761-781 (2007).
332. Andersen, J.S. & Mann, M. Organellar proteomics: turning inventories into insights. *EMBO reports* **7**, 874-879 (2006).
333. Yates, J.R., 3rd, Gilchrist, A., Howell, K.E. & Bergeron, J.J. Proteomics of organelles and large cellular structures. *Nature reviews* **6**, 702-714 (2005).
334. Au, C.E. et al. Organellar proteomics to create the cell map. *Current opinion in cell biology* **19**, 376-385 (2007).
335. Takamori, S. et al. Molecular anatomy of a trafficking organelle. *Cell* **127**, 831-846 (2006).
336. Trinkle-Mulcahy, L. & Lamond, A.I. Toward a high-resolution view of nuclear dynamics. *Science (New York, N.Y* **318**, 1402-1407 (2007).
337. Andersen, J.S. et al. Nucleolar proteome dynamics. *Nature* **433**, 77-83 (2005).
338. Rogers, L.D. & Foster, L.J. The dynamic phagosomal proteome and the contribution of the endoplasmic reticulum. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 18520-18525 (2007).
339. Ong, S.E. & Mann, M. Stable isotope labeling by amino acids in cell culture for quantitative proteomics. *Methods Mol Biol* **359**, 37-52 (2007).
340. Fasshauer, M. et al. Essential role of insulin receptor substrate 1 in differentiation of brown adipocytes. *Mol Cell Biol* **21**, 319-329 (2001).
341. Forner, F., Arriaga, E.A. & Mann, M. Mild protease treatment as a small-scale biochemical method for mitochondria purification and proteomic mapping of cytoplasm-exposed mitochondrial proteins. *J Proteome Res* **5**, 3277-3287 (2006).
342. Forner, F., Foster, L.J., Campanaro, S., Valle, G. & Mann, M. Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol Cell Proteomics* **5**, 608-619 (2006).
343. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* **39**, 1-38 (1977).
344. Fraley, C. & Raftery, A. mclust: Model-Based Clustering / Normal Mixture Modeling. (2007).
345. Team, R.D.C. R: a language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria; 2007).
346. Wilkinson, D.J. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics* **8**, 109-116 (2007).
347. Barbe, L. et al. Towards a confocal subcellular atlas of the human proteome. *Mol Cell Proteomics* (2007).
348. Ryan, M.T. & Hoogenraad, N.J. Mitochondrial-nuclear communications. *Annual review of biochemistry* **76**, 701-722 (2007).
349. O'Rourke, N.A., Meyer, T. & Chandy, G. Protein localization studies in the age of 'Omics'. *Current opinion in chemical biology* **9**, 82-87 (2005).
350. Dunkley, T.P., Watson, R., Griffin, J.L., Dupree, P. & Lilley, K.S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol Cell Proteomics* **3**, 1128-1134 (2004).
351. Ernster, L. & Schatz, G. Mitochondria: a historical review. *The Journal of cell biology* **91**, 227s-255s (1981).

352. Scheffler, I.E. A century of mitochondrial research: achievements and perspectives. *Mitochondrion* **1**, 3-31 (2001).
353. Schatz, G. The magic garden. *Annual review of biochemistry* **76**, 673-678 (2007).
354. Schapira, A.H. Mitochondrial disease. *Lancet* **368**, 70-82 (2006).
355. Detmer, S.A. & Chan, D.C. Functions and dysfunctions of mitochondrial dynamics. *Nature reviews* **8**, 870-879 (2007).
356. Vo, T.D. & Palsson, B.O. Building the power house: recent advances in mitochondrial studies through proteomics and systems biology. *American journal of physiology* **292**, C164-177 (2007).
357. Calvo, S. et al. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nature genetics* **38**, 576-582 (2006).
358. Lopez, M.F. et al. High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* **21**, 3427-3440 (2000).
359. Meisinger, C., Sickmann, A. & Pfanner, N. The mitochondrial proteome: from inventory to function. *Cell* **134**, 22-24 (2008).
360. Neupert, W. & Herrmann, J.M. Translocation of proteins into mitochondria. *Annual review of biochemistry* **76**, 723-749 (2007).
361. Peters, L.L. et al. The mouse as a model for human biology: a resource guide for complex trait analysis. *Nat Rev Genet* **8**, 58-69 (2007).
362. Ewing, R.M. et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* **3**, 89 (2007).
363. Alber, F. et al. Determining the architectures of macromolecular assemblies. *Nature* **450**, 683-694 (2007).
364. Oltvai, Z.N. & Barabasi, A.L. Systems biology. Life's complexity pyramid. *Science* **298**, 763-764 (2002).
365. Alves, R., Antunes, F. & Salvador, A. Tools for kinetic modeling of biochemical networks. *Nature biotechnology* **24**, 667-672 (2006).
366. Fisher, J. & Henzinger, T.A. Executable cell biology. *Nature biotechnology* **25**, 1239-1249 (2007).

Abbreviations

2D: two dimensional

3D: three dimensional

A2M: alpha-2-macroglobulin

AC: alternating current

ADP: Adenosine diphosphate

AGC: automatic gain control

AKT: protein kinase B, or Rac (RAC-alpha serine/threonine-protein kinase). The term AKT originates from the transformed AKR mouse strain.

AML: Acute myeloid leukemia

APC/C: Anaphase promoting complex/cyclosome

AQUA: Absolute Quantitation

ATM: Ataxia telangiectasia mutated

ATP: adenosine triphosphate

BAT: Brown adipose tissue

bis-Tris :2-[bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)propane-1,3-diol

BiNGO: Biological Networks Gene Ontology tool

BLAST, Basic Local Alignment and Search Tool

CDKs: cyclin-dependent kinases

ChIP: Chromatin innunoprecipitation

CID: collision induced fragmentation

DC: direct current

DDA: DNA damage response

DGAP: Diabetes Genome Anatomy Project

DHB: 2,5-dihydroxy benzoic acid

DMEs: drug metabolizing enzymes

DMEM: Dulbecco's modified Eagle's medium

DMSO: dimethyl sulfoxide

DNA: deoxyribonucleic acid

DNA-PK: DNA-activated protein kinase

ECD: electron capture dissociation

ECM: extracellular matrix

EDTA: ethylenediaminetetraacetic acid

EGFR: epidermal growth factor receptor

EM: Expectation Maximization

ErbB: epidermal growth factor receptor (EGFR) family

ER: Endoplasmic reticulum

ESI: electrospray ionization

EST: expressed sequence tag

ET: Evolutionary trace

ETD: electro transfer dissociation

FACS: Fluorescent-activated cell sorting

FDR: false discovery rate

FKBP: FK506-binding protein

FLP: False localization percentage

FMO: Flavin monooxygenase

FT: Fourier transform

FTICR: Fourier transform ion cyclotron resonance

FVA: Flux variability analysis

GAP: GTP activated protein

GFP: Green fluorescent protein

GLUT4: glucose transporter 4

GnRH: Gonadotropin-releasing hormone

GO: Gene Ontology

GOA: Gene Ontology Annotation

GST: glutathione S-transferase

GTP: guanosine triphosphate

HPLC: high performance liquid chromatography

ICAT: Isotope-Coded Affinity Tagging

IMAC: immobilized metal ion affinity chromatography,

IP: immunoprecipitation

IPI: International protein index

IRMPD: infrared multiphoton dissociation

iTRAQ: isobaric tag for relative and absolute quantitation

KEGG: Kyoto Encyclopedia of Genes and Genomes

LBR: Lamin B Receptor

LDL: low-density lipoprotein

LMNA: Lamin A

LMNB1: Lamin B1

LMNB2: Lamin B2

LOPIT: Localization of Organelle Proteins by Isotope Tagging

LTQ: linear quadrupole ion trap

m/z: mass to charge ratio

MALDI: matrix-assisted laser desorption/ionization

MAPU: Max Planck Unified Proteome

MBP-C: mannose-binding protein C

MCM: Mini-choromosome maintenance complex

MeCN: acetonitrile

miRNA: micro RNA

MS: mass spectrometry

MGI: Mouse Genome Informatics

mTOR: mammalian target of rapamycin

nanoESI: nanoelectrospray ionization

OB: Oligosaccharide/Oligonucleotide binding

OMIM: Online Mendelian Inheritance in Man

OXPHOS: Oxidative phosphorylation

PBS: phosphate buffered saline

PCA: Principal component analysis

PCM: polycyclodimethylsiloxane

PCP: Protein correlation profiling

PCR: polymerase chain reaction

PEP: posterior error probability

PFAM: Protein family

piRNA: Piwi-interacting RNA

PI3K: Phosphatidylinositol-4,5-bisphosphate 3-kinase

PMF: Post mitochondrial fraction

PTM: post translational modification

QCAT: concatemer of Q peptides

QIT: Quadrupole ion trap

QTOF: Quadrupole time-of-flight

Rb: retinoblastoma

RF: radio frequency

RMS: Root mean square

RNA: Ribonucleic acid

RNAi: RNA interference

ROS: reactive oxygen species

RP HPLC: reverse phase high performance liquid chromatography

RT-PCR: Reverse transcription polymerase chain reaction

SCX: strong cation exchange

SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis

SH3: Src homology 3

siRNA: Small interfering RNA

SILAC: stable isotope labeling by amino acids in cell culture

SIM: selected ion monitoring

SLD: Soft laser desorption

Smad2/3: mothers against decapentaplegic homolog 2/3

SNARE: soluble N-ethylmaleimide-sensitive factor attachment protein receptor

SOP: standard operating procedures

SORI: sustained off resonance irradiation

Src: proto-oncogene tyrosine-protein kinase Src

SULT: sulfotransferase

TCA: tricarboxylic acid

TFA: trifluoroacetic acid

TGFβ: transforming growth factor β

TGFβ R1: transforming growth factor β type 1 receptor

TMPO: Lamina-associated polypeptide 2

TOF: Time of flight

TLP: True localization percentage

TrEMBL : Translated EMBL

t-SNARE: target SNARE

UGT: UDG-glucuronosyltransferase

WAT: White adipose tissue

Y2H: Yeast two hybrid

Acknowledgements

I am extremely grateful to my thesis advisor and supervisor Prof. Dr. Matthias Mann, for giving me the opportunity to pursue my Ph.D. studies in his laboratory. The atmosphere he has created provides best of the opportunities, and a breadth of knowledge in mass spectrometry and proteomics unparalleled anywhere in the world. As a supervisor and mentor his guidance, inputs and support to me has been exceptional and very instrumental in successful completion of my projects. Throughout my studies he has been very considerate towards my personal obligations and helped me face many a tough times. As a person he is an amalgam of intelligence, inquisitiveness, compassion, and humility. His leadership qualities, personnel management abilities and scientific foresightedness have put him in the league of best scientists of the world. I am very proud to be a small part of his vision and journey towards establishing proteomics as a key discipline of the post-genomic era. I will always strive to imbibe his greatest virtues in shaping my personal and professional life.

I extend deep sense of gratitude to my previous mentors during my bioinformatics studies, especially Prof. Dr. Manju Bansal and the faculty at my alma mater - Institute for Bioinformatics and Applied Biotechnology (IBAB), Bangalore (India) for providing incomparable training in bioinformatics and allied domains. This extensive training has helped me tremendously in my current role as a researcher in proteomics and bioinformatics.

I also take this opportunity to thank Dr. Rajendra K. Bera – my mentor at IBM Software labs, Bangalore (India) for motivating me to continue with my education. I thank him for his immense belief in me and for encouraging me to pursue a career in research. His guidance throughout my tenure at IBM and even after that has helped me a lot in my current endeavors in research.

Many thanks to all the members of Department of Proteomics and Signal Transduction, for providing me with an environment, which was intellectually stimulating and conducive for interdisciplinary research I undertook during my PhD studies. Special thanks to my collaborators Dr. Jun Adachi, Dr. Cuiping Pan, Dr. Jesper V Olsen, and Dr. Francesca Forner for providing me

their excellent data, and bestowing me with their trust, which helped me to develop innovative methods for data analysis discussed in this thesis. Moreover, I will always cherish the knowledge and expertise they shared throughout these collaborations.

The contents of the thesis were reviewed and corrected by Prof. Dr. Matthias Mann, Dr. Jesper V Olsen, Dr. Chunaram Choudhary and Dr. Jürgen Cox. I thank them for taking out time from their busy schedules to review my thesis, and for providing valuable suggestions and comments for improvement. Their inputs have been very helpful in structuring the content and its presentation.

Thanks are also due to Florian Staufer and Trixi Dinkelmaier with whom I have shared my apartment for nearly three years now. They are few of the nicest people I have known, and have played a very important role in all my achievements during my PhD studies. They are my support system away from my family in India, and I truly appreciate that they were with me during my times of joys and sorrows.

I am indebted to my family for providing constant moral support and encouragement, which gave me strength to explore my capabilities and helped me tremendously to concentrate on studies and projects. My parents Shri Nand Kishore Sharma and Smt. Saraswati Devi have been the greatest source of inspiration for me and are the most important part of my existence. No words can express my feelings and gratitude for all the sacrifices they have made to provide the best upbringing to me and my brother. Thanks to my parents for imparting in me the right values, conscience, discipline and high moral standards, which have been the guiding principles throughout my life. My Ph.D. thesis and all my hard work towards its completion are dedicated to them.

Last but not the least; I thank my younger brother Pankaj Kumar for being a source of immense support and comfort. His humorous takes on pressing problems and calmness has helped me appreciate the fact that life can also be dealt with ease. He has been very supportive and prudent throughout my studies and as always a very nice brother, and my closest friend.

Curriculum Vitae

Personal Information

Name:	Chanchal Kumar
Date of Birth:	April 15, 1980
Place of Birth:	Masimpur-Assam, India
Nationality:	Indian
Gender:	Male
Marital Status:	Unmarried

Academic Qualification

2005- Till date	Ph.D. Candidate in Department of Proteomics and Signal Transduction Max Planck Institute for Biochemistry Munich, Germany
2003-2004	Post Graduate Diploma in Bioinformatics Institute of Bioinformatics and Applied Biotechnology Bangalore, India
1998-2002	Bachelor of Information Technology Acharya Narendra Dev College, University of Delhi Delhi, India

Professional Experience

- Worked as Software Engineer in Technology Incubation Center, IBM Software Labs, Bangalore, India.
  Duration: May 2004 – September 2005
  Project: Primarily worked in the domain of Bioinformatics and contributed to solutions for Health Care and Life Sciences (HCLS) sector. Also participated in IBM Academy of Technology study on "Enabling Technologies for Information based Medicine" from February 2005 - May 2005.

- Internship at IBM Software Labs, Golden Enclave, Airport Road, Bangalore, India.
  Duration: November 2003 - May 2004
  Project: Worked on the project titled "A novel method for pathway classification based on gene expression data and known gene networks". The project was in the domain of network biology and the aim was to find a novel method to classify a pathway pertaining to a specific diseased state using microarray data. I employed a methodology, which used an augmented graph theoretic algorithm for classification of pathways.

- Internship at Software Technology Parks of India, CGO Complex, New Delhi, India.
  Duration: February 2002- June 2002

Project: Worked on the project titled "E-mail tracking system". The project involved monitoring the clients mail server for the all incoming and outgoing e-mails. Subsequently an automated report on the e-mail service uses by the employees had to be prepared and stored in a database.

Publications

- Forner F, **Kumar C**, Luber CA, Klingenspor M, Mann M
  Pathway analysis of mitochondria in brown versus white adipocytes by quantitative proteomics
  (Manuscript under submission)
- Olsen JV[§], Vermeulen M[§], Santamaria A[§], **Kumar C**[§], Miller ML, Jensen LJ, Gnad F, Cox J, Jensen TS, Nigg EA, Brunak S, Mann M
  A systems view of the cell cycle by quantitative phosphoproteomics
  (Manuscript under submission)
- Pan C[§], **Kumar C**[§], Bohl S, Klingmüller U, Mann M
  Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type specific functions.
  (Manuscript accepted for publication in *Mol Cell Proteomics*)
- Bonaldi T, Straub T, Cox J, **Kumar C**, Beker PB, Mann M
  Combined Use of RNAi and Quantitative Proteomics to Study Gene Function in Drosophila.
  Mol Cell. 2008 Sept ; 31(5):762-772.
- Graumann J, Hubner NC, Kim JB, Ko K, Moser M, **Kumar C**, Cox J, Schoeler H, Mann M
  SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins.
  Mol Cell Proteomics. 2008 Apr; 7(4): 672-683
- Macek B, Gnad F, Soufi B, **Kumar C**, Olsen JV, Mijakovic I, Mann M
  Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation.
  Mol Cell Proteomics. 2008 Feb; 7(2):299-307
- Zougman A, Pilch B, Podtelejnikov A, Kiehntopf M, Schnabel C, **Kumar C**, Mann M
  Integrated Analysis of the Cerebrospinal Fluid Peptidome and Proteome.
  J Proteome Res. 2008 Jan 4; 7(1):386-399.
- Shi R, **Kumar C**, Zougman A, Zhang Y, Podtelejnikov A, Cox J, Wisniewski JR, Mann M
  Analysis of the Mouse Liver Proteome Using Advanced Mass Spectrometry.
  J Proteome Res. 2007 Aug; 6(8), 2963-72
- Adachi J[§], **Kumar C**[§], Zhang Y, Mann M
  In-depth Analysis of the Adipocyte Proteome by Mass Spectrometry and Bioinformatics.
  Mol Cell Proteomics. 2007 Jul; 6(7):1257-73.

- Macek B, Mijakovic I, Olsen JV, Gnad F, **Kumar C**, Jensen PR, Mann M
  The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis.
  Mol Cell Proteomics. 2007 Apr;6(4):697-707.
- Zhang Y, Zhang Y, Adachi J, Olsen JV, Shi R, de Souza G, Pasini E, Foster LJ, Macek B, Zougman A, **Kumar C**, Wisniewski JR, Jun W, Mann M
  MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes.
  Nucleic Acids Research. 2007 Jan;35(Database issue):D771-9.
- Olsen JV, Blagoev B, Gnad F, Macek B, **Kumar C**, Mortensen P, Mann M
  Global, In vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks.
  Cell. 2006, 127(3):635-648
- Adachi J, **Kumar C**, Zhang Y, Olsen JV, Mann M
  The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins.
  Genome Biology. 2006, 7:R80

§ Equal first author contribution

Honors and Awards

- Submitted a disclosure in IBM titled "A novel method for pathway classification based on gene expression data and known gene networks" which received "publish" rating.
- Recipient of Chief Minister's Scholarship for the Best Male Student for Postgraduate diploma in bioinformatics (2003-2004).
- Recipient of Murty's Medal of Excellence for Best Student for Postgraduate diploma in bioinformatics (2003-2004).