Development of Computational Methods for Cancer Research: Strategies for closing the feedback loop in omics workflows

Kirik, Ufuk

2015

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Development of Computational Methods for Cancer Research

Strategies for closing the feedback loop in omics workflows

Ufuk Kirik

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Engineering, Lund University, Sweden.
To be defended in Medicon Village Hörsalen, Lund,
Friday October 2nd at 09:15.

Faculty opponent
Assoc. Prof. Janne Lehtiö
Cancer Proteomics Mass Spectrometry, Department of Oncology-Pathology
Karolinska Institutet, Stockholm

LUND UNIVERSITY

Department of Immunotechnology Ideon
Medicon Village, Bld. 406, 223 81 Lund

Ufuk Kirik

Document name

Doctoral Dissertation

Date of issue

2015-10-02

Sponsoring organization

Title and subtitle : Development of computational methods for cancer research Strategies for closing the feedback loop in omics workflows

Abstract

As the ultimate workhorses of the living things, proteins undergo significant regulatory activity throughout the lifetime of a cell or an organism. Many complex diseases effect the protein composition, expression or modification in the cells or tissues they arise in. It is then no surprise that proteomics is a field full of promise, one which is expected to generate significant insights towards functional characterization of cancer and deliver potential targets of diagnostic, prognostic or therapeutic value. It is also a science in its teens, which still grows and keeps changing as it grows with the technological advances in instrumentation as the driving force.

Since data analysis routines are yet to be established fully, functional characterization of protein expression regulation in cancer remains an open question. The work presented in this thesis provides an overview of the field based on technological, computational and biological aspects in the introductory chapters, introduces a novel method for functional evaluation of changes in protein expression and demonstrates its utility in PAPER I, and describes the insights gained from investigating proteomes of several different types of human malignancies.

PAPER I underlines the challenges in functional analysis of expression data, especially from LC-MS/MS experiments, and describes a method based on a relatively simple mathematical model, which is used for subsequent analyses. PAPER II describes a study on soft-tissue sarcomas, with the proteomic analysis revealing insights to protein expression patterns potential differentiation paths. PAPER III demonstrates a pairwise comparison of malignancies of gastroesophageal track and corresponding normal tissue; we highlight several proteins and pathways as likely targets of expression regulation. PAPER IV and V on the other hand focus on breast cancer. PAPER IV presents results from an investigation of immortalized breast cancer cell lines and raises the question of how well these model systems represents the tumours they are expected to be alike. PAPER V demonstrates the therapeutic potential of inhibiting oestrogen signalling in ER+ breast cancer, using a luminal type patient-derived xenograft mouse model.

Collectively, this thesis presents some of the key concepts in quantitative proteomics workflows, elaborates on the importance of data processing routines and through the papers in the appendix demonstrates the potential of functional analysis algorithms in generating insights to cancer biology.

Key words: Quantitative proteomics, mass spectrometry, pathway analysis, bioinformatics, proteome profiling, cancer

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language. English

ISSN and key title

ISBN 978-91-7623-458-7

Recipient's notes

Number of pages

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sourcespermission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature

Date 2015-08-26

Development of Computational Methods for Cancer Research

Strategies for closing the feedback loop in omics workflows

Ufuk Kirik

LUND UNIVERSITY

Faculty of Engineering, Department of Immunotechnology

ISBN 978-91-7623-458-7
ISBN 978-91-7623-459-4 (pdf)
Printed in Sweden by Media-Tryck, Lund University
Lund 2015

Intertek
MILJÖMÄRKT 341903
fti FÖRPACKNINGS & TIDNINGS INSAMLINGEN
KLIMATKOMPENSERAT PAPPER

"Knowledge that takes you not beyond yourself is
far worse than ignorance."
Sufi teaching

from a TED talk by
Elif Safak, TEDGlobal 2010, Oxford

Contents

Original Papers

This thesis is based upon the following papers, which are referred to in the text by their Roman numerals (I-V). The papers can be found as appendices at the end of the book. PAPER I, II and IV are reprinted with permission from the publisher. Copyright 2012 (PAPER I) and 2015 (PAPER IV), American Chemical Society; Copyright 2014 (PAPER II), American Association for Cancer Research.

PAPER I
Multimodel Pathway Enrichment Methods for Functional Evaluation of Expression Regulation
Ufuk Kirik, Paolo Cifani, Ann-Sofie Albrekt, Malin Lindstedt, Anders Heyden, and Fredrik Levander
Journal of Proteome Research **2012** *11 (5)*, 2955-2967

PAPER II
Discovery-Based Protein Expression Profiling Identifies Distinct Subgroups and Pathways in Leiomyosarcomas
Ufuk Kirik, Karin Hansson, Morten Krogh, Mats Jönsson, Mef Nilbert, Peter James, and Ana Carneiro
Mol Cancer Res December **2014** 12:1729-1739

PAPER III
Quantitative protein expression analysis of gastroesophageal cancer reveals significant differences between tumors on different sides of GE junction.
Ufuk Kirik, Pehr Riessler, Peter James and Jan Johansson
(Manuscript in preparation)

PAPER IV
Molecular portrait of breast cancer derived cell lines revealed poor similarity with tumours.
Paolo Cifani, Ufuk Kirik, Sofia Waldemarson and Peter James
Journal of Proteome Research **2015** *14 (7)*, 2819-2827

PAPER V

Protein expression analysis reveals mechanisms for estrogen-independence in tumor cell subpopulations of a luminal-like breast cancer xenograft.

Nirma Skrbo, Ufuk Kirik, Alexandr Kristian, Paolo Cifani, Linn Antberg, Siver A. Moestue, Vickie Zhang, Olav Engebråten, Gunhild M. Mælandsmo, Kristin Andersen, Peter James, and Therese Sørlie

(Manuscript in review)

Author's Contribution to Papers

PAPER I
Design, development and testing of the method and software implementation, writing and preparation of the manuscript for publication.

PAPER II
Parts of the experimental design, data analysis, writing and preparation of the manuscript for publication.

PAPER III
Parts of the experimental design, data analysis, writing and preparation of the manuscript for publication.

PAPER IV
Carried out the data analysis (together with PC), particularly the functional analysis of the results, contributed to the manuscript.

PAPER V
Contributed to experiment design, carried out the data analysis (together with NS), particularly the functional analysis of the results, contributed to the manuscript.

Excluded Publications

Protein Expression Changes in Ovarian Cancer during the Transition from Benign to Malignant
Sofia Waldemarson, Morten Krogh, Ayodele Alaiya, Ufuk Kirik, Kjell Schedvins, Gert Auer, Karin M. Hansson, Reto Ossola, Ruedi Aebersold, Hookeun Lee, Johan Malmström, and Peter James
Journal of Proteome Research **2012** *11 (5)*, 2876-2889

Radio-resistance does not correlate with DNA-repair enzymes concentration in breast cancer cell lines but with anti-apoptotic pathways: A comparative study
Paolo Cifani, Ufuk Kirik, Louise Jonsson, Linn Antberg, Karin Hansson and Peter James
(Manuscript in review)

Abbreviations

BCCL	Breast cancer cell line
BRCA1/2	Breast cancer type 1/2 susceptibility protein
BioPAX	Biological Pathways Exchange
ChiBE	Chisio BioPAX Editor
CV	Coefficient of variation
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DIGE	Difference gel electrophoresis
ECD	Electron capture dissociation
ECM	Extracellular matrix
EGFR	Epidermal growth factor receptor
ER	Estrogen receptor
ES	Enrichment score
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FCS	Functional class scoring
FDR	False-discovery rate
FEvER	Functional evaluation of expression regulation
FT-ICR	Fourier transform ion cyclotron resonance
GA	Genetic algorithm
GO	Gene Ontology
GSEA	Gene set enrichment analysis
HCD	High-energy collision dissociation
HER	Human epidermal growth factor receptor
HGP	Human genome project
hMSC	Human mesenchymal stem cell
HPLC	High-performance liquid chromatography
HR MAS MRS	High-resolution magic angle spinning magnetic resonance
HTML	Hyper-text markup language
ICAT	Isotope-coded affinity tag
IPA	Ingenuity Pathway Analysis
IT	Ion-trap

iTRAQ	Isobaric tag for relative and absolute quantitation
kNN	k-nearest neighbours
LC	Liquid chromatography
LMS	Leiomyosarcoma
LTQ	Linear trap Quadrupole
m/z	mass/charge ratio
MALDI	Matrix-assisted laser desorption ionization
MES	Maximum enrichment score
MFH	Malignant fibrous histiocytoma
MRM	Multiple reaction monitoring
MS	Mass spectrometer/spectrometry
MS/MS	Tandem mass spectrometry
ORA	Over-representation analysis
PARP	Poly ADP ribose polymerase
PCA	Principal component analysis
PDX	Patient derived xenografts
PR	Progesterone receptor
PSM	Peptide-spectrum match
PTM	Post-translational modification
ROI	Region of Interest
RP	Reverse phase
SBML	Systems biology markup language
SC	Spectral counting
SCX	Strong cation exchange
SDS	Sodium dodecyl sulphate
SILAC	Stable isotope labelled amino acids in cell-culture
SRM	Selected reaction monitoring
STS	Soft-tissue sarcoma
SWATH	Sequential window acquisition of all theoretical spectra
TCA	Tricarboxylic acid
TMT	Tandem mass tagging
TNM	Tumour-size, node, metastasis
TOF	Time-of-flight
TSQ	Triple stage quadrupole
UPS	Undifferentiated pleomorphic sarcoma
VEGF	Vascular endothelial growth factor
XIC	Extracted ion chromatogram

Summary:
The Biological Clockwork

What makes tumour cells tick? Imagine the potential for personalized medicine if only we could understand, in detail, how tumours arise and develop. Those days might not be that far away...

The -omics era

The human body is often compared with machines of varying types with different regulatory and control mechanisms, which one can model and try to understand. Several branches of modern biomedical research target different levels of this general machinery. One such branch is *genomics*, that is the study of how the genetic code at the core of an organism is structured and how it is used, or regulated, over the lifetime of the organism in question. The completion of the Human Genome Project (HGP) allowed the scientific community to have a blueprint of the molecular building blocks of our bodies. The surprising result is that almost our entire genome, that is the protein coding regions of our DNA, is shared not only among all living humans, but also an overwhelming portion is shared among all mammalians.

That being the case, the questions arise as to what mechanisms account for all the visible and invisible differences between us humans. The answer lies in the next level of biological molecules, the proteins. Just because a gene sequence is shared between two human beings, it does not mean that they express the same protein, or that the particular protein in question is functional to the same extent for these two individuals. *Proteomics* is yet another, somewhat newer, branch of biomedical research which focuses on qualitative and quantitative study of the proteome, that is the whole set of proteins in a living system such as a cell, a tissue sample or an entire organism. Proteomics studies the identity, quantity or function of proteins that exist within the studied system. While the genome of an organism is maintained stabile over its life span, its proteome is highly dynamic. In other words the proteome is subject to significant regulatory activity depending on both “time” and “location”, continuously throughout the lifetime of the organism.

To illustrate this point, consider a nerve cell in your eye and a cell on your skin. They look and work completely differently as they express different proteins, yet they have largely the same set of genes. Similarly, different proteins can be found in the same cell during different phases of its replication cycle. Another such example is the change in protein expression of cells that are exposed to environmental factors such as nutritional variations or different kinds of stress such as lack of oxygen or ionizing radiation.

Role of proteomics in cancer research

The overarching goal in many proteomics projects is to analyse the changes in the proteins between two or more conditions, especially diseases like cancer where large-scale changes in the proteome occur. A deeper understanding of these changes will ultimately help researchers develop better medicine, as well as clinicians tailor more effective treatments for the patients.

The majority of traditional cancer medication is composed of toxic molecules effecting mainly dividing cells. Therapeutic approaches are largely based on the assumption that malignant cells divide faster and more often than "normal" cells in their surrounding. This approach is essentially the medical counterpart to carpet-bombing in a battlefield. This rather grim and tragic analogy is mostly due to the collateral damage associated with the approach. Systemic cytotoxic medicine causes serious side effects to the patient, which has a major impact on quality of life. To make the matters worse, in some cases patients benefit relatively little from the treatment, if any at all, due to acquired resistance to therapy.

At the same time, the economic burden of over-treating patients is not negligible. Not only does conventional therapy little good to the patient, but it also puts significant burden on the health-care systems, due to primary and secondary effects of conventional chemotherapy. Many patients have compromised immune systems and are susceptible to opportunistic diseases.

Personalized medicine

These points are all common arguments for personalized medicine. The concept refers to tailoring the treatment of a cancer patient based on the specific form of disease the patient has. That way the patient receives the therapy that has the highest chance to giving health benefits, systemic cytotoxic agents are used to a minimum and the patient hopefully experiences minimal impact on quality of life. However, in order for personalized medicine to become a global reality certain conditions need to be met. First of all, in order to tailor the treatment to a patient, the mechanisms affected by the disease need to be discovered. Pro-oncogenic mutations, that is the changes in the genome that promote tumour development or survival need to be identified. Networks of interacting proteins, often called pathways, which have been altered in the malignant cells, need to be analysed in detail. These are tasks that are far from being trivial.

While whole-scale personalised cancer therapy is not here yet, there is significant progress towards the end goal. Today in Sweden almost every cancer patient leaves samples that are investigated with state-of-the-art technology, revealing many important insights at different levels of the mechanisms; key mutations are highlighted, pathways are investigated with genome-/proteome-wide discovery studies. In some cases significant improvements are observed.

One such success story is the use of Herceptin in HER2-positive breast cancer patients. HER2 is a receptor protein that sits on the cell membrane, and this protein is typically involved in signal transduction that is important for cell survival and development in the tumours. Herceptin is a molecule that inhibits the signalling process through this receptor protein and thereby hinders the development of the cells that express this protein highly. Another promising example is the use of small molecules called tyrosine-kinase inhibitors (TKI), particularly in patients that have a specific type of cancer in the abdomen, called gastrointestinal stromal tumours (GIST).

In both of the cases above, the treatment is aimed at targeting a protein of interest, typically a weak-point in a pathway that is susceptible to inhibition. Based on these principles, the work presented in this thesis is a collection of studies aimed at improving our understanding of the proteomes of different types cancer. In PAPER I, a novel computational method for evaluating expression regulation is presented. This method, and the software that comes with it, provides preclinical researches the means to identify the mechanisms by which tumour cells differ from their normal counterparts.

The remaining four papers present studies focused more on data analysis and biological interpretation in different settings. In papers II and III, clinical samples from soft-tissue sarcomas and gastroesophageal tumours are analysed, respectively, and new insights to these complex diseases with key proteins and pathways are highlighted.

The studies presented in papers IV and V are focused on breast cancer biology. In PAPER IV, an overall poor correlation in protein expression of immortalized cell lines and the tumour subtypes that they are presumed to represent was shown. This finding points towards the fact that better care must be taken when transferring knowledge gained by studies model systems. In PAPER V, the presented study focuses on the therapeutic aspects of ER-positive luminal-type breast cancer, specifically the systemic significance of oestrogen signalling pathway, and the therapeutic potential of inhibition of this pathway is shown.

While it would be too optimistic to give a timeframe for curing all forms of cancer, the field of cancer therapy will certainly see major improvements in survival rate and patient quality of life over the next 15-20 years. As our understanding of these diseases increase through systems biology, more sophisticated treatment options that target multiple proteins in a pathway, or multiple pathways, will emerge and personalised medicine will be more of a reality rather than a concept out of science fiction books.

Foreword

After approximately 20 years and $3bn, one of the most significant research projects in human biology, the Human Genome Project (HGP) was completed in 2003 to provide a blueprint of the molecular building blocks of our bodies. Arguably the largest landmark achievement since the discovery of the double-helix structure of DNA, the completion of HGP not only answered many questions but also opened up new horizons to explore for researchers within biomedical sciences.

Research into genomics in the past decade clearly demonstrated that two cells sharing a common gene does not imply that both cells express exactly the same gene products or that a particular protein product of the gene in question is functional to the same extend in these two cells. The number and importance of the questions that could not be explained with the insights into the genome alone have no doubt paved the way for studies aimed at proteins and regulatory mechanisms involving proteins. The mechanisms involved in DNA replication, transcription and translation were shown and explained relatively early on following the discovery of the structure of the DNA molecule (CRICK 1958). However much like in the classic novel, the rabbit hole was subsequently shown to go much deeper than originally thought.

From protein expression to cancer therapeutics, the research that gave birth to this thesis is of cross-disciplinary nature and is built on three main pillars, which cover different aspect of data generation, analysis and interpretation of this data in biological context. These three subjects are presented in consequential order in the following chapters, followed by an appendix containing five papers presenting the research in the context of this thesis.

Proteomics

> "A few years ago the idea of making proteins or polymers "fly" by electrospray ionization (ESI) seemed as improbable as a flying elephant, but today it is a standard part of modern mass spectrometers."
>
> John B. Fenn, Nobel lecture, 2003

Definition, origin and history of proteomics

Proteomics, refers to the qualitative and quantitative study of the *proteome*, that is the entire set of proteins, expressed by a genome at a certain time under specified conditions. The term proteome was coined in 1994 by Wilkins during a conference in Siena, Italy and started appearing in literature since then (Wasinger et al. 1995; Wilkins et al. 1996).

The motivation to study the proteome stems from the fundamental observation that the proteins are the primary actuators both within and outside the cells of a system. They function as enzymes in metabolic and signalling events, as structural components of the cytoskeleton and of virtually all organelles, as transport molecules as well as regulatory elements by which protein and gene expression is controlled (Raven 2005). In that sense one could crudely summarize that genomics studies what may happen in a cell, whereas proteomics study what actually happens.

There are several complications associated with the study of proteins in a biological system, however. For instance while the genome of an organism is maintained as stable as possible, the proteome of an organism is highly dynamic. In other words the proteome is subject to significant regulatory activity in both spatial and temporal domain, continuously throughout the lifetime of the organism. Considering the functional and physiological differences between the wide variety cells, which share virtually the same genome, present in the body of an individual, the role of regulatory mechanisms involving proteins becomes self-evident. Similarly, the protein expression profile, that is which proteins are expressed and to what degree, of a cell in phases of the cell cycle differs significantly. Thus any discovery study of the proteome, aiming to map out the proteins in the target system, is essentially a snapshot of the proteome at a given time and at a specific location.

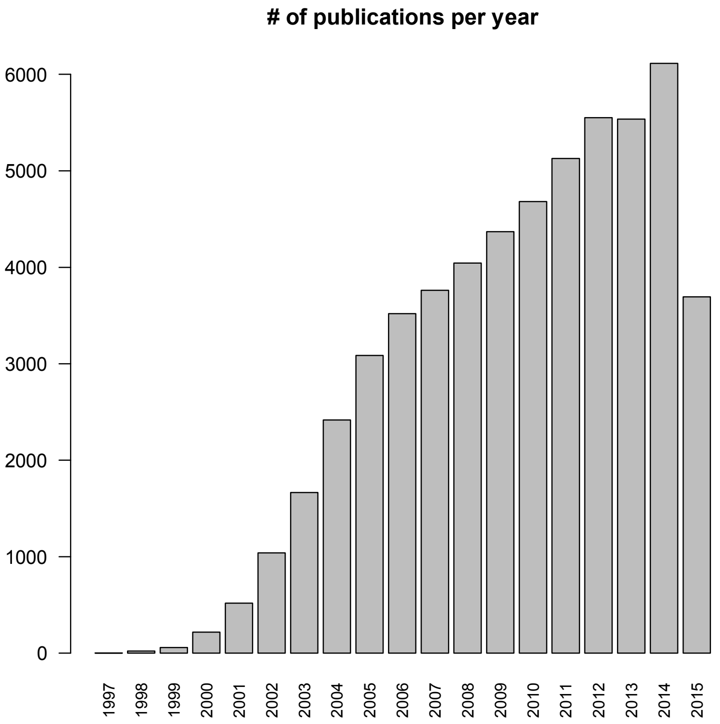


Figure 2: The rise of proteomics in literature; number or articles on PubMed with the term "proteomics" in title or abstract as of July 2015.

Another major challenge in studying the proteome originates from the difference in levels of abundance associated with different proteins. Specifically, proteins are observed to be expressed between approximately 7 and 10 orders of magnitude in cells(Beck et al. 2011) and plasma(N. L. Anderson and Anderson 2002), respectively. Operating on such a range requires analytical methods that can cover a wide dynamic range, i.e. equally adept at detecting or measuring proteins that are expressed from few copies up to billions per cell.

Proteins abundances span such a large range as a consequence of the many functions that proteins fulfil in living organisms. Low abundance proteins are typically characterized by high turnover rates, and may have drastic effects on the regular functioning of the system. Examples of such proteins could be initiator kinases for signalling cascades, where the initial signal is amplified several folds downstream until the signal reaches the target(s). Another example could be detector proteins that sense stress conditions like hypoxia, or damage to vital components of the cell such as the DNA. High abundance proteins, on the other hand, are characterized by specific fundamental functions that are required in high numbers; examples of such are cytoskeletal proteins or ribosomal proteins that typically turn over at a slower rate.

The diversity of proteins occurs not only on the level of abundance (copy number), but also on physical properties of these macromolecules. Specifically, proteins show tremendous variability with respect to length, and thus molecular weight (MW), ranging from a handful of amino acids (*Tuftsin*, 501 Da) up to well over 30 000 amino acids (*Titin*, over 3.8 MDa), according to the April 2014 release of UniProtKB (UniProt Consortium 2015). Historically, a divide and conquer approach has been utilized for examination of proteins, where protein(s) of interest are cut into shorter chains of amino acids.

One of the first methods for investigating peptides, called Edman degradation (also referred to as Edman sequencing) named after the Swedish chemist Pehr Edman who described the procedure in 1950 (Edman 1950), is based on a divide-and-conquer approach as well. The procedure, involves repetitive steps of cleaving the N-terminal amino acid residues from the peptide backbone, resulting in a single amino acid and a peptide sequence that is one shorter than the original. The identity of the single amino acid can be determined with an orthogonal method such as chromatography or electrophoresis, while the process can be iterated to cleave the next residue on the N-terminus. This method is slow and laborious, especially on a large scale where for instance the proteome of a whole cell lysate is studied. Furthermore, the utility of the method is limited to cases where the N-terminal amino acid is accessible and not chemically modified, as well as when working with intact proteins. There have been variants of sequential sequencing, which allow for C-terminus degradation in a similar fashion (Stark 1968), however with the developments in mass spectrometry instrumentation and ionization techniques in the later decades of 20th century and onwards, this method is practically obsolete for most proteomic studies.

In the current years, what one could refer as the modern era of proteomics, two orthogonal methods have become the norm: *affinity proteomics* and *mass-spectrometry proteomics*. The studies presented in this thesis will focus exclusively on the latter.

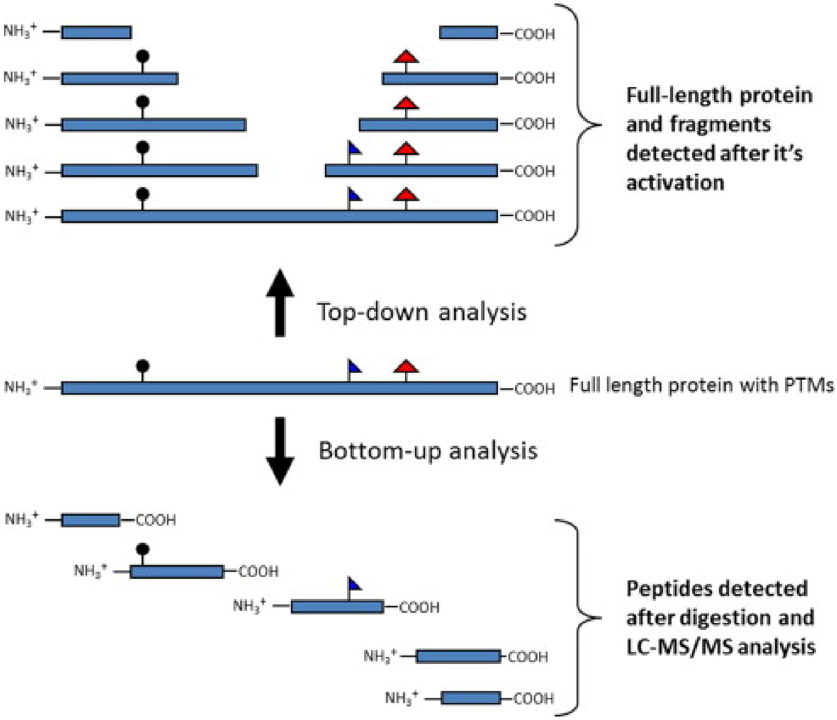
Mass spectrometry proteomics

A mass spectrometer (MS) is an analytical instrument that measures the mass over charge ratios (*m/z*) of ionized analytes in gas phase. It is essentially composed of an ion source, a mass analyser to measure the *m/z* values and a detector to count the ion intensity (Aebersold and Mann 2003).

A typical MS-proteomics experiment follows a 4-step workflow, which starts with sample preparation, separation of proteins or peptides and acquisition of mass spectra and finally data analysis. There are essentially two main approaches in MS-proteomics, often referred to as the *top-down* and *bottom-up* mass-spectrometry. In both cases analytes are measured, fragmented and measured again, in a process referred to as tandem mass-spectrometry, or often in the short form MS/MS. In this context the spectra obtained in the first scan is called MS^1 while the secondary scan(s) following the fragmentation are referred to as MS^2. This technique can be extended to include further fragmentation, sometimes referred to as multi-stage mass-spectrometry or MS^n, mainly through the use of an ion-trap instrument.

Top-down vs. bottom-up MS/MS

The main difference between these two approaches is in the analytes that are subjected to mass spectrometric analysis. In top-down proteomics, intact proteins are analysed, whereas in bottom-up proteomics, proteins are digested into peptides prior to injection to the spectrometer.



NH_3^+
COOH
Full-length protein and fragments detected after it's activation
Top-down analysis
Full length protein with PTMs
Bottom-up analysis
Peptides detected after digestion and LC-MS/MS analysis

Figure 3: Top-down vs. bottom-up proteomics, reproduced from (Scherl 2015) with permission from the publisher.

The most common approach to a MS-proteomics experiment is the bottom-up approach, where the proteins are digested into peptides, often using a protease, during the sample preparation step, thus increasing the sample complexity by approximately two orders of magnitude.

The masses of the intact peptides are measured in MS1, then fragmented often using a collision chamber filled with an inert gas, and the resulting fragments are measured in MS2. These fragmentation spectra are matched to libraries of mass-spectra, either of theoretical mass-spectra created by *in silico* digest of the target proteome or so-called spectral libraries containing experimental MS2 spectra of that particular peptide. This approach is sometimes called *shotgun proteomics*, analogous to shotgun genome sequencing. One of the major advantages of bottom-up proteomics is better possibilities for peptide separation, in one or more dimensions, prior to mass-spectrometry analysis compared to intact proteins, which makes this method better suited for analysis of complex samples where proteome coverage is of main concern (Yates, Ruse, and Nakorchevsky 2009).

On the downside, one of the fundamental drawbacks of bottom-up proteomics is the breaking of protein-peptide linkage (Figure 4). Once digested, it is impossible to tell which protein a particular peptide comes from, unless it is unique to a particular protein. This is a rather complicated issue that gives rise to a multitude of problems such as the protein identification problem and the protein quantification problem(Nesvizhskii and Aebersold 2005; Claassen 2012). Since the bottom-up approach relies on identification and quantification on peptide-level, with the protein-peptide linkage gone, it becomes challenging to figure out which proteins were in the sample originally. One method that has been popular, mostly due to its simplicity, is the so-called *Occam's razor*, where the idea is to explain the evidence with the simplest answer. In MS-proteomics context, it translates to describing the observed set of peptides with the fewest number of proteins. For unique peptides, the situation is simple as there is enough evidence for including the proteins containing these peptides in the resultant dataset. The situation is far more complicated however for so-called degenerate peptides, i.e. peptides that could originate from multiple different proteins. This situation becomes more complex for protein isoforms and splice variants, often giving rise to protein groups, rather than distinct proteins. While a wide range of different statistical methods are developed to tackle the protein identification, also known as the protein inference, problem (Serang and Noble 2012), a dominant design is yet to emerge and a considerable amount of groups still use Occam's razor principle in their workflows.

Another consequence of breaking the protein-peptide linkage is the difficulty in the quantification of proteins. In the simplest case, where a protein is identified by two unique peptides with intensities I_1 and I_2, the determination of the abundance of this protein becomes a function of these two intensities. This too is a non-trivial task, as ionization efficiency and thus signal yield varies significantly between peptides. Similar to the protein inference problem, no dominant design has emerged despite numerous different methods developed over the recent years. The most commonly used methods are often simple alternatives such as taking the mean or median of values from individual peptides, or summing the intensities and taking that sum as a proxy for relative protein abundance (Carrillo et al. 2010). The use of stable isotope labelled standards spiked in at known concentrations helps in this regard, allowing for better comparisons, however the underlying problem of re-constructing the relative protein abundance remains one of the major challenges for bottom-up proteomics.

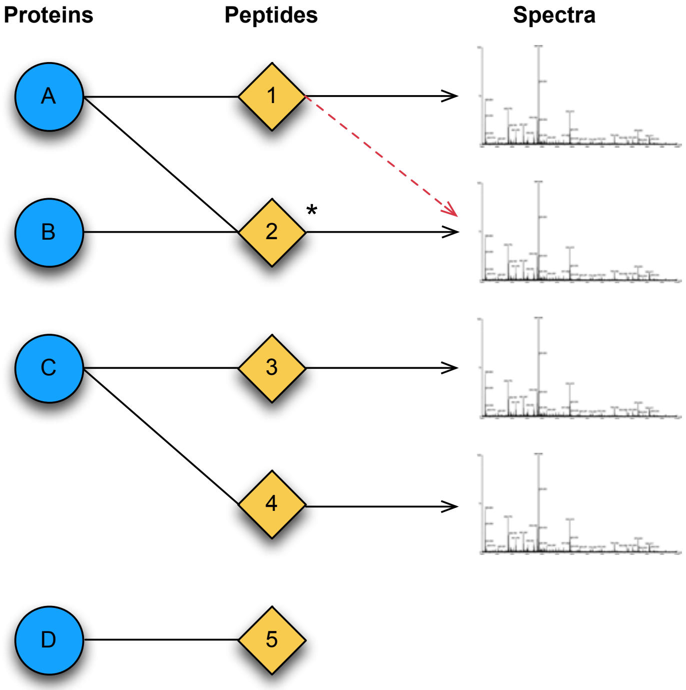


Figure 4: Protein-peptide linkage issue in bottom-up proteomics. A majority of tryptic peptides will not be unique to a single protein, which raises questions as to how to identify which proteins were in the sample originally, and in what abundance. Here peptide 2 is a *degenerate* peptide, in other words, whether it originates from protein A or protein B cannot be determined certainly. This fact not only jeopardizes the identification of protein B, but also casts doubt upon the quantification of protein A and B. In this case, the identification and inclusion of protein B is entirely dependent on the strategy employed for identification procedure, also called as the protein assembly problem.
Protein quantification is also non-trivial for bottom-up proteomics; here protein C has two unique peptides, 3 and 4, thus the abundance of protein C needs to be determined as a function of the abundances of these two peptides. Finally since the complexity of the sample is increased approximate by two orders of magnitude by protein digestion, it is likely that a number of peptides will not be identified at all, and in the grossly simplified case above, protein D will be neither identified nor quantified in the sample.
Another complication that occurs in this scheme is management of the so-called *chimeric* spectra, which contain fragmentation patterns of multiple co-eluting peptides within the same isolation window (denoted with the red dashed arrow). Based on the acquisition method these spectra may not be matched to any theoretical spectra during the identification of peptides.

In comparison, the top-down proteomics approach relies on the measurement of intact proteins and their fragments, usually created by electron capture dissociation (ECD) (Zubarev et al. 2000) or electron transfer dissociation (ETD) (Syka et al. 2004). The direct advantages of the top-down approach are the higher sequence coverage of the analysed proteins as well as better characterization of post-translational modifications that may otherwise be lost in the process (Kelleher et al. 1999). The top-down approach also avoids inference problems that arise with breaking of the protein-peptide linkage. In practical terms this means both the identification and quantification at protein level can be done with much higher accuracy.

Despite the advantages, top-down proteomics have not been standard approach for proteomics experiments due to difficulties associated with front-end separation of proteins. This is issue becomes even more challenging when separating post-translationally modified copies of proteins from the unmodified ones. A range of separation techniques, both off-line and on-line, exist for intact proteins(Capriotti et al. 2011; Catherman, Skinner, and Kelleher 2014), each with their own potential advantages and drawbacks. For instance, while off-line separation techniques might result in better separation, they increase sample handling and the throughput decreases, rendering large-scale experiments impractical.

A secondary issue is the identification of the analytes observed in the mass-spectrometer. Bottom-up analysis often relies on matching of MS^2 spectra to theoretical spectra for peptides resulting from *in silico* protein digestion. However fragmentation of intact proteins is much more complex and computational methods for identifying proteins in top-down experiments are rather limited in comparison (Calligaris, Villard, and Lafitte 2011; Catherman, Skinner, and Kelleher 2014).

The choice between top-down and bottom-up approaches boils down to a compromise between sequence coverage and proteome coverage, as well as sample complexity. Both approaches have specific advantages and shortcomings, and thus will likely co-exist and evolve side-by-side in the foreseeable future. Since proteome coverage has been the priority for the works presented in this thesis, all five papers included in this thesis have adopted bottom-up approach.

Post-translational modifications and implications

Post-HGP studies have shown that the complexity of the proteome vastly supersedes that of the genome. In humans, a genome of approximately 20 000 genes, is estimated to give rise to over 1 million distinct proteins (Jensen 2004), thanks to a series of pre- and post-translational processes. Mechanisms such as genomic recombination, differential transcription initiation (at alternative promoters) and termination and alternative splicing generate different mRNA transcripts from a single gene, altogether adding approximately an order of magnitude to the complexity (Pan et al. 2008). An additional order of magnitude is attributed to processes that proteins undergo after they have been synthesized, commonly referred as post-translational modifications (PTMs).

While the exact number and function of all possible PTMs are up for debate and beyond the scope of this thesis, more than 200 modifications have been shown in the literature (Walsh 2006). It is clear that they play a critical role in cellular machinery as they often alter the structure and thus the function of the modified protein. To further complicate things, many PTMs are highly dynamic and are subjected to change due to stimuli. A particular type of modification may occur at various different positions on a single protein in a non-exclusive manner and different modifications may occur in tandem on the same protein and some modifications on a specific residue may take shape in many different ways, e.g. glycosylation.

Given the importance of the PTMs, many groups have spent decades studying different types of modifications. However as discussed earlier, the shotgun (bottom-up) approach to proteomics have some limitations with respect to studying these modifications. One of the most studied PTMs (Khoury, Baliban, and Floudas 2011), phosphorylation, typically require high-level of enrichment of the modified peptides prior to shotgun LC-MS/MS analysis. Not only does enrichment procedure imply more labour in sample preparation and increased risk of sample handling variability but also the separation of modified peptides from the unmodified ones obscure information regarding *site occupancy*. While it's possible to carry out two LC-MS/MS runs, one for the phospho-enriched sample and one for the flow-through direct comparison of quantification is still far from trivial due to losses in connection with enrichment procedure.

The focus on post-translational modifications, in particular phosphorylation, in the recent years have resulted in hundreds of thousands of reported phospho-sites, however only a few of these have been manually validated by orthogonal methods, and yet even fewer have been characterized functionally, thus giving rise to what can be referred to as the phosphoproteome interpretation gap.

Sample preparation and labelling

The exact steps taken during the sample separation are tightly connected with the chosen experimental workflow. Labelling strategies, protein digestion (for bottom-up approaches) and separation steps are typical procedures prior to MS analysis.

While different proteases can be used for this purpose, trypsin has been the most common choice due to its specificity, optimal working conditions as well as the product peptide lengths. Trypsin normally cleaves on the C-side of the peptide backbone after arginine (R) or lysine (K), thus having two advantages: i) based on the relative occurrence of these amino acids, tryptic peptides are often between 10 and 20 amino acids long, which is ideal for fragmentation inside the mass spectrometer, and ii) since both these residues are basic they tend to pick up charge relatively easy and thus “fly well” in the magnetic/electric fields inside a mass-spectrometer.

Labelling strategies

Labelling of samples allows for multiplexing of runs and helps control experimental variability. Essentially sample labelling comes in two flavours, metabolic and chemical labelling.

Metabolic labelling is based on introduction of stable-isotope labels into the cells as they grow and replicate. The general idea is that while the chemical properties of stable-isotope labelled peptides or proteins do not differ significantly from the natural counterparts, the instruments will detect a mass shift for the labelled analytes. The most commonly used form of metabolic labelling is *stable-isotope labelling with amino acids in cell culture* (SILAC) (Ong et al. 2002), which is based on the incorporation of isotopically labelled versions of amino acids. Arginine and lysine with heavier nitrogen and/or carbon isotopes that are 10Da and 6Da, heavier than their natural counterparts respectively are often used.

Another type of metabolic labelling is based on incorporation of heavy nitrogen (^{15}N) instead of ^{14}N into the growth medium. This approach can be used to label single-celled microorganisms, and even higher organisms, while SILAC approach is primarily used for cultured cells (Gouw, Krijgsveld, and Heck 2010). The SILAC approach however has a fundamental advantage in that the amount of labels per analyte is known *a priori*, since trypsin cleaves after a lysine or arginine. Furthermore, SILAC labelling has been utilized in several different approaches over the years. One such approach is the pulsed-SILAC (pSILAC) strategy whereby effects of any treatments can be observed in protein turnover rates (Schwanhäusser et al. 2009) or the use of a heavy-labelled internal standard for comparison of multiple SILAC experiments, where the samples of interest are unlabelled, an approach called Super-SILAC (Geiger et al. 2010).

SILAC approach has been utilized in several of the papers in this thesis. In Paper IV, we have used SILAC labelling for one of the cell lines, which was subsequently used as an internal standard across multiple comparisons with other cell lines. In Paper V, the same setup was utilized for as an internal standard comparing tumour samples. Additionally, one of the datasets from the Super-SILAC study by Geiger et al. has been used as a stress test for the algorithm in Paper I.

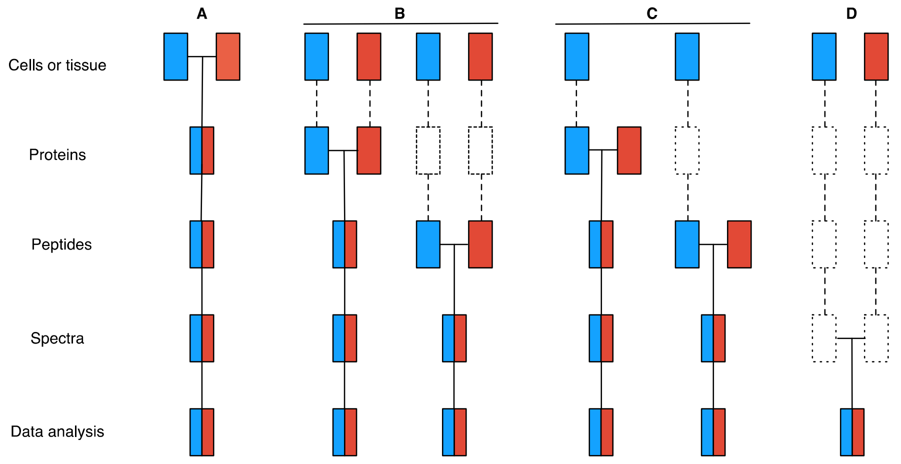


Figure 5: Different workflows used in quantiative proteomics with respect to sample preparation and labeling; colored boxes represent different conditions, horizontal lines depict the stage at which the samples are combined. Steps marked with dashed lines are susceptible for experimental variability. Metabolic labeling (**A**) is the best alternative in minimizing experimental variability, while chemical labeling (**B**) and spike-in standards (**C**) at protein or peptide level still provide some means for control of variability. In label-free analysis (**D**) samples are run individually and are thus susceptible to variability, or errors, during the whole workflow. Quality control mechanisms are only possible once in silico, once the results are collected and combined during the data analysis step. (Adapted from (Bantscheff et al. 2012))

Metabolic labelling is an effective method to minimize experimental variability and quantification errors, especially since it can be introduced at the beginning of the workflow. However, as the name suggests, this approach requires labels to be introduced to the samples metabolically, which prevents the use of these methods in human samples *in vivo* due to both ethical and practical reasons.

Chemical labelling of proteins, or peptides, after their extraction circumvents this problem and allows its application for human samples. Isotope-based chemical labelling was introduced by *Gygi et al.* with the ICAT approach, which is based on deuterium tagging of cysteine residues (Gygi et al. 1999). In the following years chemical labelling gained momentum with the introduction of isobaric labels. The concept of isobaric labelling is to covalently bind the target peptides, or proteins, with reagent molecules that have the same mass. After fragmentation, the reagents leave residues of differential masses, called reporter ions, which are then used for quantification. The two most popular flavours of isobaric labelling reagents are iTRAQ(Ross et al. 2004) and TMT (Thompson et al. 2003), the former providing 4- or 8-plex, and the latter providing up to 10-plex throughput (Werner et al. 2014).

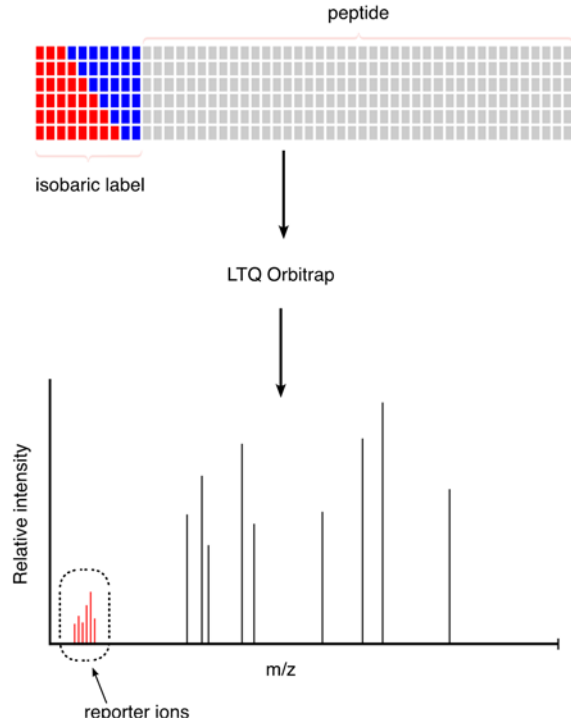


Figure 6: Conceptual representation of TMT labeling used in papers II and III. A number of tags that have the same total mass but different chemical structures are attached to peptides of interest (upper panel). Since the precursor ion mass is equal for all peptides, they are selected for fragmentation together. After the fragmentation, the relative abundance information is acquired by comparing the reporter ions (from the HCD chamber) while the identification information is read-out from the remaining portion of MS2.

In a label-free approach, the samples are run individually and results are combined and compared *in silico* after the experiment. This approach has the advantage of being fastest and cheapest, as it entails minimal sample preparation and no expensive reagents for used for labelling. However, this approach is also susceptible to the highest amount of variability as the samples are handled individually, as well as due to the stochastic nature of the sampling in the mass spectrometer. It is thus imperative to have the necessary quality assurance mechanisms in place for reliable data analysis (addressed in detail by colleagues at our department (Sandin et al. 2014; Chawade et al. 2015)). Since most experiments aim at comparing protein expression levels in a number of samples, the lack of overlap due to under-sampling becomes an important issue that needs to be tackled. There are several reasons why a protein might be identified in one run and not in another, and in fact, it is a rather complicated problem (Aebersold 2009). One study estimates that about 16% of the peptide features were targeted for MS/MS and only half of those eventually identified (Michalski, Cox, and Mann 2011). One common strategy to mitigate this problem is by matching features that is peaks corresponding to a single peptide ion, between different label-free runs. This way if a feature is selected for MS/MS and a peptide-spectrum match (PSM) is found, then this identification can be "propagated" across other samples where there is a matching feature, but no identification.

It should be noted that these workflows can be combined to utilize the advantages and to avoid the drawbacks of each specific approach. In fact, with the exception of paper I, in which an algorithm is presented, studies presented in this thesis all feature multiple workflows combined into one big investigative narrative. Specifically, in papers II and III, where we have set out to investigate the proteomes of soft-tissue sarcomas and gastroesophageal tumours, respectively, the investigation starts with two dimensional, differential in-gel electrophoresis (2D-DIGE). Following this exploratory analysis, samples were analysed using 6-plex TMT labelling, with one of the six channels dedicated for a internal standard obtained from pooling all samples. Based on the findings from TMT-studies we have re-formulated our question(s) and carried out label-free analysis of individual tumours where we got the best proteome coverage.

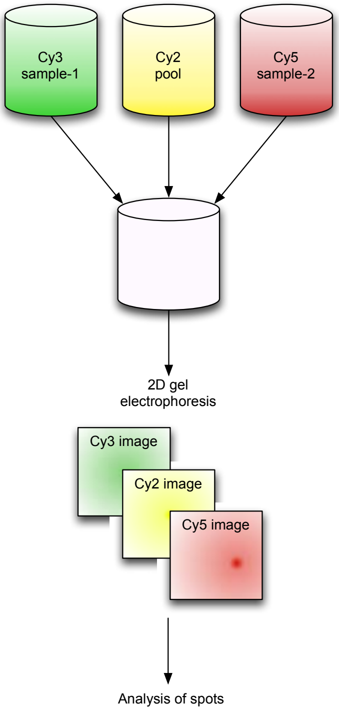


Figure 7: Conceptual representation of 2D-DIGE analysis. Two samples of interest are labeled together with a standard, with three fluorescent dyes and mixed. Analytes are separated in two orthogonal dimensions (often mW and pI) and analyzed in an optical scanner for differential expression patterns.

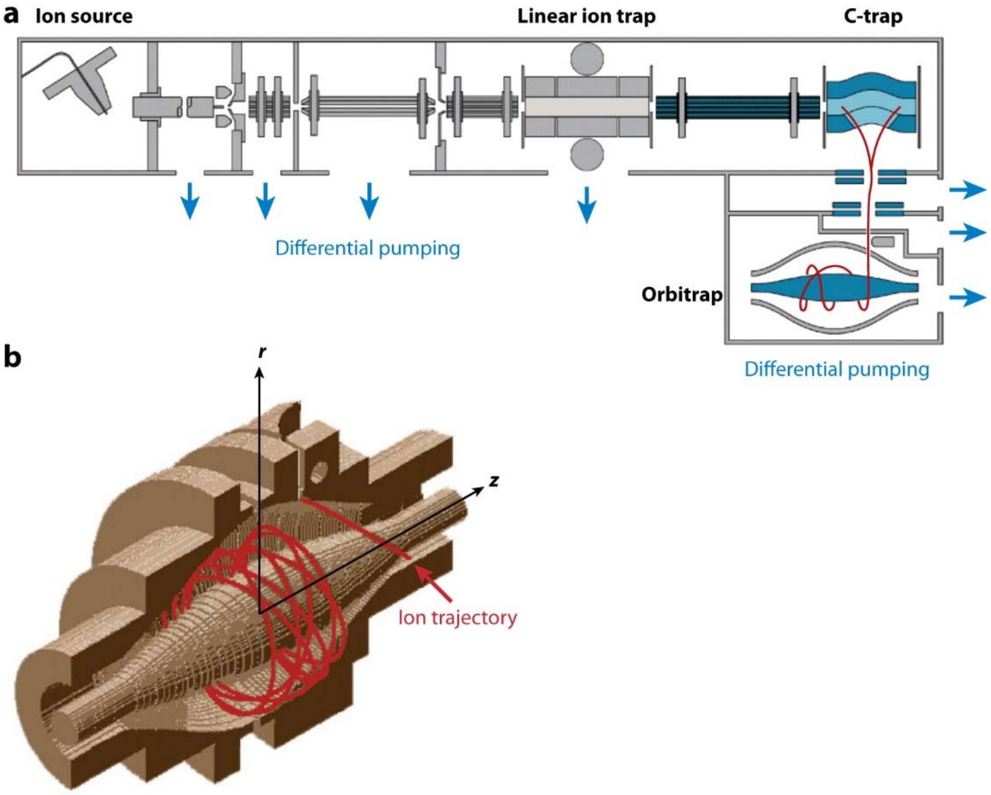
Different types of MS instruments

As mentioned earlier mass spectrometers have three main components; an ion source, one or more mass analysers and an ion detector. Different types of components have certain advantages and drawbacks, and thus are fit for different types of analyses (Aebersold and Mann 2003; Domon and Aebersold 2006; Yates, Ruse, and Nakorchevsky 2009).

Mass spectrometers measure the *m/z* ratio of ions, and thus high-throughput analysis of proteomes require efficient ionization techniques, which had been a challenge until the emergency of two so called soft-ionization methods; matrix-assisted laser desorption ionization (MALDI) (Tanaka et al. 1988; Karas et al. 1987) and electrospray injection (ESI) (Fenn et al. 1989). While the former is based on a pulsed laser and has the potential to give insights into the spatial distribution of proteins in tissue slices (Thiele et al. 2014), ESI has become popular in MS-proteomics due to its continuous nature and for the possibility for coupling to a chromatography outlet, as well as production of multiply charged ions (Aebersold and Mann 2003; Domon and Aebersold 2006).

Mass analysers come in many different flavours, however it is possible to cluster them into two broad categories; scanning analysers and trapping analysers (Yates, Ruse, and Nakorchevsky 2009). A time-of-flight (TOF) analyser calculates the *m/z* values for the ions based on their flight time, while a quadrupole (Q) instrument uses oscillating radio frequencies to selectively permit the passage of ions of a specific *m/z* value. These two types of devices can be used in tandem to create hybrid MS/MS setups, such as Q-TOF, TOF-TOF and QQQ (also called triple stage quadrupole TSQ) (Aebersold and Mann 2003; Domon and Aebersold 2006). Mass analysers that are based on the idea of trapping ions include the ion trap (IT), Fourier transform ion cyclotron resonance (FT-ICR) and Orbitrap spectrometers. These too, can be run in parallel, typically with a quadrupole device, one of which is the hybrid LTQ-Orbitrap instrument (Makarov, Denisov, Lange, et al. 2006; Makarov, Denisov, Kholomeev, et al. 2006) (Figure 8) which provides the advantages of both devices. The high-resolution and mass accuracy from the Orbitrap is typically utilized for MS^1, while the speed and sensitivity of the linear trap quadrupole (LTQ) is used for MS^2 acquisition per precursor ion (Yates, Ruse, and Nakorchevsky 2009; Eliuk and Makarov 2015).

The mass-spectrometric analyses carried out for the studies included this thesis were done using an ESI-LTQ-Orbitrap XL instrument interfaced with an Eksigent nanoLC+ HPLC system.



Yates JR, et al. 2009.
Annu. Rev. Biomed. Eng. 11:49–79

Figure 8: Schematics of the LTQ-Orbitrap instrument (**a**) and cross-sectional schematic of the Orbitrap analyzer (**b**), reproduced from (Yates, Ruse, and Nakorchevsky 2009) with permission from the publisher.

Acquisition methods

Proteomics is a field with lots of different workflows and data acquisition strategies during tandem mass-spectrometry analysis are no exception. There are three main strategies used in the field (Figure 9), which differ from each other based on the choice of precursor ions for fragmentation and detection.

In data-dependent acquisition (DDA), also referred to as *shotgun proteomics*, the instrument is set to select ions detected in MS^1 for fragmentation, and subsequent MS^2 acquisition, during which time the first mass analyser is occupied with the next survey scan in an Orbitrap instrument. The exact number of ions that can be selected for fragmentation is dependent on the instrument used. This strategy requires no prior knowledge regarding the analytes in the sample. The quantitation could be done using the MS^1 or MS^2 spectra, depending partially on the labelling approach. For instance, for isobaric labelling the quantitation is carried out in MS^2, using the reporter ions. In SILAC experiments, however, the quantitation is carried out by extracting the precursor ion chromatograms (XIC) and integrating the area under the curve. For label-free experiments, both MS^1 and MS^2 could be used and there are many different approaches described in the literature (Bantscheff et al. 2012; Schulze and Usadel 2010).

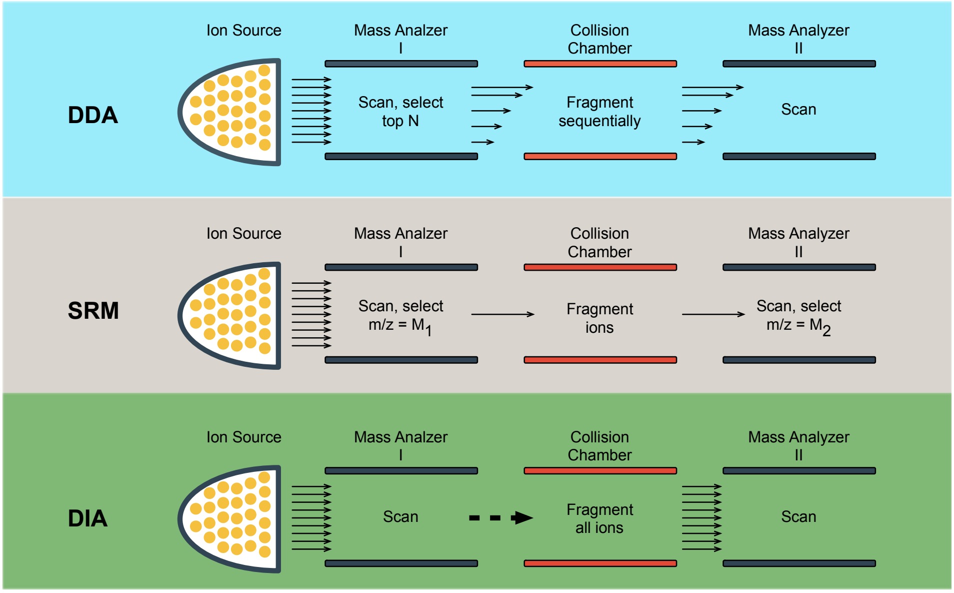


Figure 9: Simplified representation of different acquisition modes in tandem mass-spectrometry. In data-dependent acquisition (DDA), the instrument selects the top N most intense ions for fragmentation; fragments ions are measured in the second mass analyzer. Selected reaction monitoring (SRM) approach, the instrument is set to pick ions that match a specific m/z value at a given retention time in MS^1, and measure only specific fragment ions in MS^2. Data-independent acquisition (DIA) is based on the idea to decouple the selection of ions for fragmentation from the MS1 intensities, occurring either in sequential or multiplexed random mass windows.

Shotgun proteomics can be likened to taking a snapshot of the proteome at a given time-point for all cells in the sample. The prevailing assumption in the field is that while individual cells might have differing proteomes at a particular time point, averaging over the whole sample will even out the outliers and yield a meaningful average. One potential pitfall with this approach is that those outliers might indeed be the most important ones biologically. Especially considering the stochastic selection of ions for fragmentation the DDA approach has significant drawbacks in terms of reproducibility and dynamic range.

An alternative approach, called single reaction monitoring (SRM), addresses these drawbacks by only selecting certain ions with a specific m/z value at a given time. Often referred to as *targeted proteomics*, an analyte can be very accurately and reproducibly measured given a retention time window, precursor ion *m/z* and daughter ion *m/z* values. These three pieces of information is often called an *assay* and is specified prior to the analysis. While there are many quantification strategies, this approach generally provides the most reliable quantification and the lowest limit-of-detection amongst the mass spectrometric methods (Picotti and Aebersold 2012). There are however several shortcomings of this approach such as the limitations with respect to the number of proteins that can be measured in a single run, the necessity of *a priori* information for assays and the quality thereof. The SRM approach has been reported to yield robust quantification of proteins given high quality assays (reviewed in (Hüttenhain et al. 2009; Picotti and Aebersold 2012; Bantscheff et al. 2012)) however this method, by definition, will not yield any information besides the targeted proteins, thus leading to no new insights regarding protein composition of the sample. It is also limited by the availability of unique peptides for a protein that was discussed earlier with the Occam's razor approach to bottom-up proteomics.

Recent improvements in instrumentation technology have given rise to a third alternative data-independent acquisition (DIA). While the proof-of-principle DIA concept has been described over a decade ago (Purvine et al. 2003; Venable et al. 2004), the practical application and broader adoption of DIA approach is relatively new. While the precise implementations vary, the basic idea of DIA is to allow a number of precursors go in to fragmentation simultaneously, regardless of their MS1 intensity, usually based on wider isolation windows (see (Chapman, Goodlett, and Masselon 2014; Bilbao et al. 2015) for in-depth reviews). The DIA approach provides a compromise alternative between the DDA and SRM approaches, aiming to retain the discovery aspects while avoiding the under-sampling problem. These advantages come at the cost of complexity however, as the MS^2 spectra from DIA experiments are significantly larger and more complex than those originating from DDA experiments. This problem presents an important challenge in terms of data analysis in proteomics workflows.

The studies presented in this thesis are of discovery nature, aiming at characterizing the proteomic changes in several different conditions, and thus have adopted a DDA approach. As our experimental designs often included consolidation and comparison of multiple runs, especially in papers II and III, the prevalence of missing values for proteins, due to either stochastic sampling or false identifications, across the whole dataset has been a concern. We alleviated this problem using *in silico* chromatographic alignment and subsequent propagation of peptide identifications on peptide level, as well as a well-known imputation algorithm at the protein level (in Paper III).

While this approach is sufficient to yield some insights into the proteomes of the samples of interest, more robust and accurate comparisons of proteins require SRM-based experiments, given a set of interesting target proteins can be identified. For discovery proteomics, two different paths of development exist. The first relies on technical advances related to mass-spectrometers resulting in increased sampling rates. While this is a likely development, considering the evolution of DDA datasets it is unlikely that the technical advances will dramatically improve the detection of low-abundance proteins in complex samples. The second path of development is associated with improvement of data analysis routines associated with the DIA approach. In an ideal scenario, we should be able to get full proteome coverage if all co-fragmenting peptides can be efficiently de-convoluted and identified. It is likely that both DDA and DIA methods will co-exist and co-evolve in the coming years.

Regardless of the choice of acquisition the results will ultimately depend on a series of data processing steps from spectra to biological interpretation, which will be explained in the following chapter.

Bioinformatics approaches

“Essentially, all models are wrong, but some are useful.”
George E. P. Box, *Empirical Model Building and Response Surfaces,* 1987.

“Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.”
John Tukey, The Future of Data Analysis, 1962.

The overall scope of the studies presented in this thesis, as well as the work done by the author is in the cross-disciplinary domain between biology, mathematics, statistics and computer science, often referred to as *bioinformatics*.

Bioinformatics as a field aims to develop tools to further the understanding of biological data. The term bioinformatics has been overused, however, becoming a roof under which all and any non-wet-lab biology fits, and the bioinformatician then becomes anyone who predominantly works *in silico*.

Proteomics, much like its older sibling genomics, aims at and relies on high-throughput, large-scale datasets. Analysis of such datasets pose a series of challenges, from the redundant and efficient storage of data to inferential data mining; all of which falls within the scope of bioinformatics. As mentioned earlier, the works in this thesis are based on MS-proteomics and analysis of shotgun MS-data, and thus this chapter will cover the aspects related to better and more efficient analysis of MS-data, specifically on a functional level.

Steps from the instrument to an answer

A typical experiment follows the relatively simple workflow (Figure 10), originating from a question of interest. Given the question, an experimental design is devised, the laboratory work for sample preparation is done and the samples are made ready for injection into the instrument. While the alternative approaches in sample preparation and instrumentation are numerous, time-consuming and labour-intensive, these steps are still often straightforward and typically follow well-documented methods.

The steps following data acquisition are neither as well documented nor are the options as clear. The data typically goes through a pipeline of software tools used for identification, quantification, identity propagation, quality assurance measures and normalization. At every step along this pipeline, some level of statistical inference is done and a level of uncertainty is introduced. The overall goal however is to gain some new insight and find an answer to the original question. In an ideal scenario, once a new insight or a new piece of the puzzle is found, new questions arise, closing the feedback loop in the workflow. In MS-proteomics these last steps in data analysis typically correspond to some form of functional analysis; that is inference regarding the implications of the observed information on the proteome. Paper I presents a novel method that is designed to facilitate functional analysis of shotgun LC-MS/MS data, by extrapolating the notion of expression regulation from protein level to pathway level and evaluating pathways based on this regulation model.

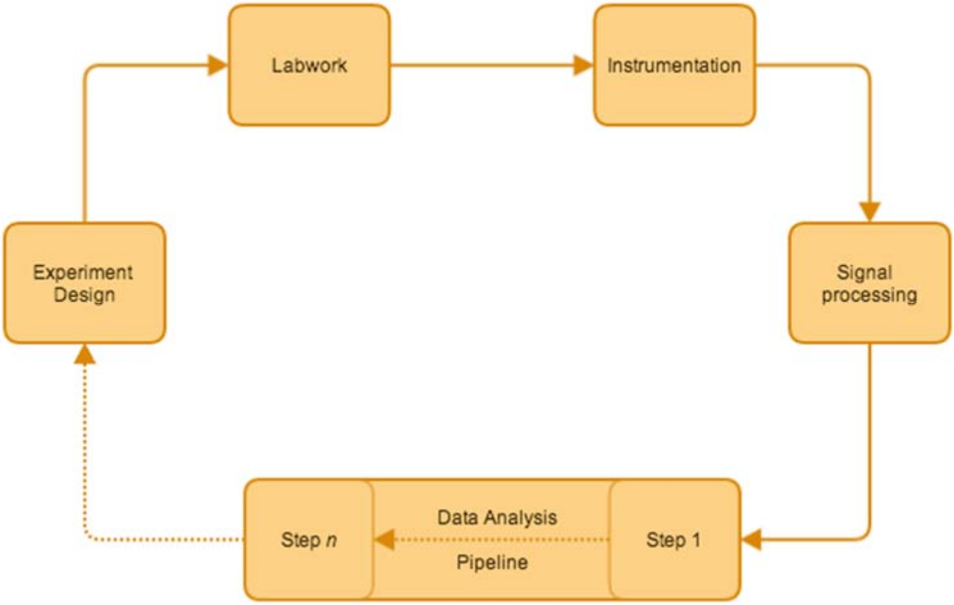


Figure 10: The typical workflow employed in MS-proteomics experiments. The scope of the work presented in this thesis, and the focus of the author, falls within the last steps of data analysis, and the feedback loop to experiment design through hypothesis generation.

Data processing

Once the data is collected by the mass spectrometer, it typically goes through a series of data analysis steps, which have a significant impact on the final results of the experiment. Some of the most common steps in typical MS data analysis pipelines include: spectrum matching and peptide identification, peptide quantification, protein assembly/inference, protein quantification and normalization.

Unless an SRM approach is adopted, where target proteins are predetermined, the identities of the proteins in the sample need to be determined. This is usually accomplished by identification of fragmentation spectra of the peptides in the sample and *in silico* assembly of protein composition, the latter earlier discussed within the comparison of bottom-up and top-down proteomics approaches. Identification of peptides from mass spectra can be accomplished in two ways: *de novo* sequencing (Ma et al. 2003) or matching fragmentation spectra to a spectral database (Eng, McCormack, and Yates 1994).

Spectral databases typically contain *in silico* digestion and the fragmentation of the target proteome, based on heuristics approaches. There are numerous different search algorithms (Deutsch 2011) each with its own set of parameters, which leads to emergence of local best practices and further complicates reproducibility of results across labs. Furthermore it has been shown that it is possible to combine results from multiple search engines, increasing the number of identifications given a false-discovery threshold (Shteynberg et al. 2013). An alternative approach to database searching is to match experimental MS/MS spectra to a library of empirical spectra, ideally from the same type of instrument. As of July 2015, The National Institute of Standards and Technology (NIST) peptide tandem mass spectra library, one of the largest spectral libraries, contains over 3.8 million spectra of more than 1.2 million entities.

Following database searching and spectral matching protein assembly and quantification are often the next steps in the pipeline. Problems associated with identification and quantification of proteins were introduced and discussed in the previous chapter. While normally independent from each other, there are promising approaches to address both problems together (Forshed et al. 2011; Webb-Robertson et al. 2014). Furthermore, normalisation of quantitative values is often needed to correct for experimentally introduced differences. A novel tool, called *Normalyzer*, was developed at our lab to evaluate different normalization algorithms for omics data (Chawade, Alexandersson, and Levander 2014).

The end product of the above mentioned pipeline is typically a list of proteins measured in one or more conditions. Especially in discovery phase shotgun experiments, this long list of proteins with associated intensities, or ratios in case of SILAC or isobaric labelling, rarely answers the original biological questions, but rather the goal is often to do some inference on biology based on the protein expression data. Statistical methods such as principal component analysis (PCA) and hierarchical clustering are often used methods in this context for investigation of any underlying patterns in the data.

We have used these methods in studies explained in Papers II – IV. In Paper II, we identify differential expression patterns among tumour samples we had in the experiment, and based on the difference we propose the existence of different subtypes among these tumours. In Paper III, we investigate whether or not protein expression patterns differ between two closely related groups of tumours. Lastly, in Paper IV hierarchical clustering is used as a measure of similarity across different types of cell lines. It should be noted however, that these methods are sensitive to missing values in the data, which may be solved by either by excluding proteins that have not been measured in many samples, or by imputing the missing values. An imputation algorithm called k-nearest neighbours (kNN) is used for clustering in the study described in Paper III.

While these statistical methods may help reveal underlying patterns in the expression data, it is often desirable to continue the analysis process by investigating the functional implications of the observed results.

Statistical enrichment analysis

Much like in genomics, the end product of MS-proteomics is a long list of analytes measured under one or more conditions. Given the ultimate goal to extract new knowledge out of experiments, the common approach is to group the genes/proteins in the dataset into functional sets such as Gene Ontology (GO) terms, interaction networks, pathways etc. The main appeal of this approach is that identification of active networks or regulated pathways between different conditions have more explanatory power than a long list of proteins with corresponding expression values (Glazko and Emmert-Streib 2009). The explanatory power of higher-level functional sets has been demonstrated by Mootha and colleagues in a study where, after multiple-hypothesis correction, no single gene was found to be differentially expressed between type II diabetes positive and negative individuals. However when looking for sets of genes, the researchers could identify a set of genes associated with oxidative phosphorylation that were differentially regulated in human diabetic muscle (Mootha et al. 2003).

An overwhelming majority of the enrichment methods that have been developed in the past decade consider these functional sets as *bags of genes,* and thus can trivially be generalized to proteins. These methods can be divided into two categories: tools that carry out over-representation analysis (ORA) and what Pavlidis and colleagues refer to as functional class scoring (FCS) methods(Pavlidis et al. 2004). ORA-based methods calculate the probability of observing a certain number of genes from a functional set, given a list of genes of interest and a background dataset. In practical terms, this type of analysis aims to answer questions of the type: *"What's the probability that out of 100 genes in our target list, 25 are a member of GO term GO:0006412 by pure chance?"* There are two main consequences to ORA methodology: first, the definition of the "target list" will play a major role on the results of the analysis, and second, even though the target list may be defined as genes that are differentially regulated above a certain threshold the ORA-based methods do not use the expression values at all but rather tests functional sets for over-representation, often using a hyper-geometric distribution.

FCS approach aims at using the experimental information better, specifically by ranking the genes in dataset according to a function that uses the available experimental information. As a second step, an enrichment score for each functional set is calculated based on the ranked list of genes and lastly the enrichment score of each functional set is tested for significance. FCS-based methods differ from each other in the specific implementation of ranking of genes, the set-level statistic and assessment of significance. Specifics and performance of the FCS-based methods have been reviewed extensively over the years (Khatri, Sirota, and Butte 2012; Maciejewski 2014; Laukens, Naulaerts, and Berghe 2015; Nam and Kim 2008; Fridley, Jenkins, and Biernacka 2010; Emmert-Streib and Glazko 2011). Specifically the type of significance testing in this context, competitive or self-contained (Goeman and Buhlmann 2007), has been debated at length (Nam and Kim 2008; Khatri, Sirota, and Butte 2012; Maciejewski 2014), however it has been pointed out that these two types of tests are not objectively comparable in terms of statistical power and that they test different aspects of the data (Emmert-Streib and Glazko 2011). This issue has been discussed in Paper I in the context of pathway analysis.

Pathway analysis and enrichment modelling: The FEvER model

Given MS-based proteomics data, the task of identifying pathways and networks that have been subject to systemic regulation have several conceptual and practical challenges that need to be tackled.

One such challenge is completeness of datasets; in comparison to transcriptomics experiments, datasets in MS-based proteomics are still rather incomplete since only a portion of the proteome can be quantitatively measured. This is partially due to the dynamic range problem associated with stochastic sampling, as mentioned previously. Furthermore, while the sampling process is biased toward highly abundant proteins, it has been shown that each replicate analysis will not necessarily sample the same portion of the proteome and thus complicates the further analysis of the results (Picotti, Aebersold, and Domon 2007; Wolf-Yadlin et al. 2007; Gstaiger and Aebersold 2009; Picotti et al. 2009; Tabb et al. 2010). However, due to temporal and local dynamics of the proteome, it is rather difficult to estimate how large of a portion of the proteome is accessible through mass-spectrometry proteomics. In other words, given the disparity between approximately 5000 proteins in a hypothetical dataset and 20500 annotated genes in the human genome, it is not trivial to estimate the portion of the "true proteome" that has been quantified. This is due to the fact that not all genes are transcribed in all cells, at all times. The true size of the proteome in a complex sample, like human tissue samples, will likely be a matter of debates for several years to come. Additionally, having adopted a bottom-up approach distinctive quantification of proteins from detected peptides is often not possible. Instead measured intensities are typically associated with groups of proteins that cannot be distinguished from one another given the observed peptides. Due to these properties of MS-based proteomics data, methods devised within the transcriptomics field are not necessarily suitable for MS-based proteomics data out-of-the-box. Besides the technical challenges there are conceptual challenges that stem from the complicated relationship between proteins and pathways.

To address these challenges, we have developed a model for evaluation of functional regulation, based on expression data. This model relies on three fundamental assumptions/heuristics and a concept, which we refer to as *region of interest* (ROI). ROI is defined as a portion of the dataset that is considered to be statistically and biologically significant enough to be considered more than random fluctuations, and is set by the scientist running the experiment. The assumptions/heuristics are described as follows:

ROI presence: *It is statistically unlikely that a high number of proteins in a pathway are significantly regulated by pure chance.* This follows from fundamental combinatorics; given a set of elements where a few of the elements are different (in this context, different translated to *regulated significantly*) the probability of finding a certain number of these in a subset follows hyper-geometric distribution, which can in turn be approximated with binomial distribution, given that the set size is much larger than number of significant elements.

As discussed earlier in this thesis, data coming from mass-spectrometry proteomics is rather incomplete and not all proteins in the sample are identified and quantified. Considering that fact, the probability of observing "*one more significant protein*" for a particular pathway decreases rapidly for each significant protein in a pathway. In other words, with each additional significantly regulated protein, the significance of regulatory activity on the pathway is higher.

Functional specificity: *Pathways identified with proteins that do not have many pathway associations have less ambiguity, and thus are more significant biologically.* Proteins typically have many different functions and majority of the proteins that have pathway annotations are associated with multiple pathways. This leads to what can be referred to as an ambiguity in pathway inference, similar to the protein inference problem discussed earlier. Since the goal is to do inference on pathway level expression regulation, given a protein $\boldsymbol{p}$, the certainty of inference decreases as the number of pathways which $\boldsymbol{p}$ is a part of increases.

Co-regulation: *Systemic regulation of a pathway is correlated with the cumulative expression regulation of individual proteins in that pathway.* Regulation of activity through a particular pathway is a complex matter that has multiple components, such as modification or compartmentalization of proteins, or changes in abundances of non-organic analytes such as metallic ions. However, it is rather straightforward to postulate that cumulative expression regulation of proteins in a pathway will be contributing factor of regulatory activity on that pathway.

We have developed and tested functions to evaluate pathways based on these three heuristics and calculate an enrichment score (ES) as a weighed linear combination, such that for each pathway

$$S_1 = f\text{(proportion of proteins in ROI)}$$

$$S_2 = -\log\text{ (normalized specificity in ROI)}$$

$$S_3 = \text{normalized cumulative exp. reg.}$$

$$ES = \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3$$

where *f* refers to a function approximating the factorial function. While the details of the functions have evolved over time, the principal ideas of the model have remained the same.

This experimental enrichment score is tested for significance using mock data based on the same dataset, in a Monte Carlo manner. The underlying idea is that in a biologically meaningful dataset, regulatory activity is not randomly distributed to genes or proteins and there will be co-variation patterns, which should not exist in a mock dataset. Thus the significance of observed enrichment score of a particular pathway is estimated by calculating the proportion of mock dataset with which the same pathway scores as high, or higher given the enrichment model.

Randomization

The creation of mock data, or *randomization*, can be done in several ways, however since no new identifications can be made, the mock datasets should contain the identical set of proteins as the real experimental dataset. In Paper I, we discuss three different approaches to randomization in this context; permutation of ratios, sampling from a lognormal distribution and sampling from an estimate empirical distribution.

Permutation is a commonly used, non-parametric method for randomization of values or labels. However, given there are several thousand entities in a typical dataset, the number of possible permutations is astronomic. To illustrate this point, the whole set of permutations for a dataset consisting of 2000 proteins is approximately 3.3×10^{5735}, which is beyond comparison with anything earthly. In this context it is unclear how many permutations would be necessary in order to get a representative figure, or if it would be feasible to implement that many permutations in practice. Another issue with permutation is based on the fact that the same values will be used in all mock datasets, which multiplies the effect of the outliers.

One alternative to permutation is to take random samples from the underlying distribution, in this case that of the expression values for proteins in the dataset. However, the precise underlying distribution is unknown and thus needs to be estimated. Biological values are often considered to be normally distributed, owing to the central limit theorem (McDonaldUniversity of Delaware 2009). Given ratios of intensities for proteins in two different conditions, logarithms of the ratios typically show normal-like distribution. Thus one option for generation of mock values could be achieved by sampling from a lognormal distribution with appropriate parameters such as $\mu = 0$ and standard deviation approximated with that of the experimental data.

The assumption of normality (of log ratios) is not necessarily supported, and should be investigated explicitly for each dataset. There are various ways for testing normality however it is not trivial to implement a fully automated test to incorporate into a computational workflow. Furthermore, the choice of distribution parameters is likely to play a role in the sampling process. The proposed solution to this issue is to avoid having assumptions about the underlying distribution but instead model the distribution empirically. This is done by a method called the variable kernel method with Gaussian smoothing (Silverman 1986), as implemented in Apache Commons Math library.

Implementation

The original implementation of the method was developed using Java 6 SE programing language, allowing for cross-platform, multi-threaded execution. The tool assumes a rather straight-forward workflow (Figure 11): the user submits the data, the data is parsed and the necessary canonical information is queried using HTTP requests to Pathway Commons (Cerami et al. 2011) online tool. This particular resource is a collection point of multiple different pathway information databases.

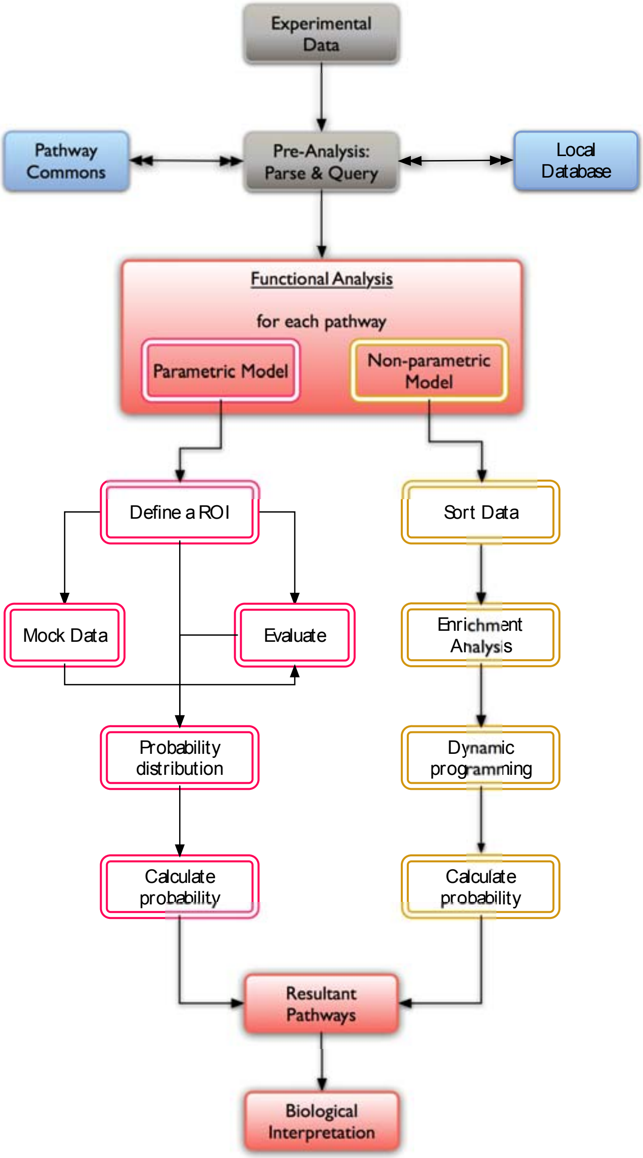


Figure 11: Workflow diagram of the FEvER software

For comparative reasons a non-parametric enrichment model based on the popular GSEA method (Subramanian et al. 2005), with modifications to the significance calculation based on the dynamic programming approach described by Keller et al. in GeneTrail (Keller, Backes, and Lenhof 2007; Backes et al. 2007) was implemented. This approach uses a running sum test statistic as the enrichment score and takes the maximum deviation from 0 as the maximum enrichment score (MES). The exact probability of observing that particular MES_{obs} is then calculated by considering all possible paths the run-sum statistic could have taken, with corresponding MES values (Figure 12).

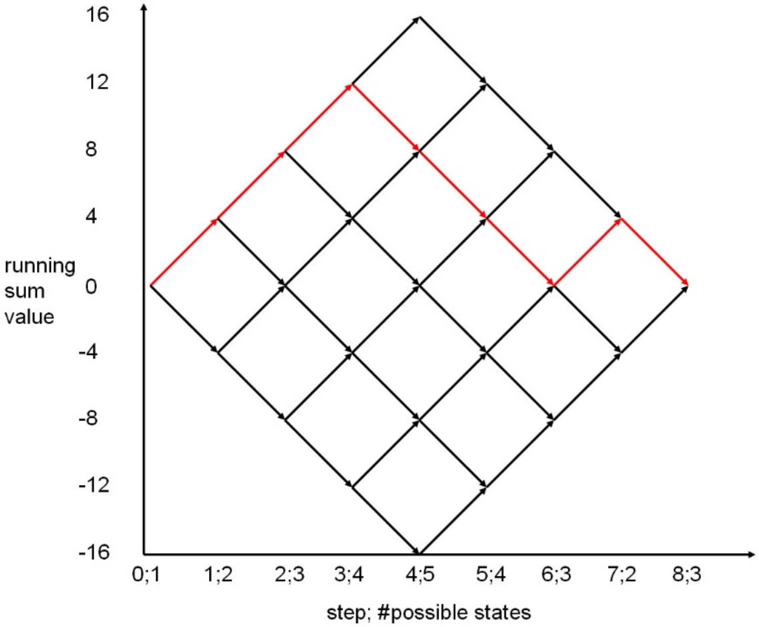


Figure 12: Schematic visualization of the running sum statistic. All possible running sum statistics for an ordered list of 8 genes of which 4 belong to a functional group (i.e. pathway) are shown. The procedure iterates over the ordered list and the statistic is increased, or decreased based on whether or not current gene is a part of the pathway under evaluation. The red labeled running sum statistic has a MES value of 12 and the corresponding p-value is 1–54/70=0.229. Th e numbers on the x-axis refer to the index and the number of possible running sum values in the current step (Reproduced from (Keller, Backes, and Lenhof 2007) with permission from the publisher).

The two enrichment models then get combined and presented back to the user in a series of HTML-based reports. While having two different scores for each pathway might be unconventional or counter-intuitive initially, the complementary statistical approaches and the strengths of the different methods makes multi-model evaluation worthwhile. A consensus between different models indicates significance of estimated expression regulation for a particular pathway, regardless of the statistical model used. The two enrichment models have different tendencies in identifying likely regulated pathways. Contradicting significance scores typically indicate a small pathway with a couple of highly regulated proteins (high parametric, low non-parametric score) or a widespread low-level differential expression over a larger pathway (low parametric, high nonparametric score). The latter case could be a symptom of a too-strict ROI for the data set, but could also indicate a systematic bias between the samples compared, for instance if slightly fewer cells or sub-optimal digestion in one sample compared to the other one.

Nevertheless, a consensus score was implemented as a combination of the significance scores reported from the two models. This consensus score, called the META score, is implemented as follows

$$S_{META} = \left(\frac{\log_{10}(PAR)}{\log_{10}(PAR_{max})}\right)^{c_1} \cdot \left(\frac{\log_{10}(NPAR)}{\log_{10}(NPAR_{max})}\right)^{c_2} \cdot 100$$

and ranges between from 0 to 100 for increasing significance, where a META score of 100 means that the pathway is flagged as extremely interesting by both models. The META score is used for quick comparison and sorting of pathways.

Model validation and stability analysis

The validation of functional analysis results is a non-trivial task as pathways are abstract constructs that cannot be measured and physically observed. Furthermore, the cumulative knowledgebase on pathway regulation is still relatively rudimentary and thus it is difficult to talk about false positives and false negatives. A "false discovery" could very well be a yet unknown discovery or effects of a poorly understood cross talk. For that reason the final validation of any pathway analysis result should be a biological test of hypothesis developed based on the results.

One alternative is to look for true positives, i.e. regulatory activity on well-studied pathway, in relatively simple experimental setups. In paper I, we analysed at proteomes of yeast grown in glucose or ethanol rich medium, looking for metabolic processes that would reflect the regulatory differences in the two samples. As expected, we observed extremely high scores from *TCA cycle, aerobic respiration, electron chain transport* and *glyoxylate cycle* pathways, with both enrichment models. This finding is in well agreement with previous results from similar studies (Kolkman et al. 2005; Futcher et al. 1999). Furthermore, the parametric model we developed highlighted *gluconeogenesis* pathway, which has not gotten a significant score ($1.176 \text{ x } 10^{-1}$) from the non-parametric model. Given the rechanneling of carbons in the ethanol-rich medium through reverse glycolysis (Futcher et al. 1999), this pathway is indeed relevant in the comparison of the two yeast populations.

We presented two more analyses in paper I, where we have investigated the functional changes given a dataset of proteomic changes in ductal versus lobular breast cancer samples (Geiger et al. 2010) as well as a scalability test based on a transcriptomics study of a human cell line (MUTZ-3) resembling dendritic cells exposed to a sensitizing chemical 1-chloro- 2,4-dinitrobenzene (DNCB), taken from a larger study carried out in our lab (H. Johansson et al. 2011). In both cases, the method has proven useful and identified relevant and interest pathways that warrant in-depth studies.

As described in the Proteomics chapter, data originating from MS proteomics experiments are susceptible to variability in both identification and quantification. In order to investigate how robust the method is to variation in experimental data, we have introduced perturbations to a dataset at varying levels and measured Pearson's correlation between reported pathway scores from the perturbed dataset and the original version of the dataset. Two different stability tests were carried out, to address variability in quantification and identification respectively (explained in supplementary material to Paper I).

To test for stability against variability in quantification of proteins, we have introduced noise in the form of a standard Gaussian random variable added to the log2-ratios in the dataset. The second test was devised to investigate the effect of missing values on the dataset. Missing values were introduced to the dataset by randomly removing quantification values for 1, 5 and 10% of the proteins in the dataset. As a quantitative tool FEvER excludes proteins without quantification values from the analysis and treats these proteins as if they have not been identified.

Table 1: Correlation between scores from perturbed and original datasets, with varying levels of perturbation. Pearson's R2 values are given in the table.

Level of perturbation	**Noise Test**		**Missing value test**	
	PAR	NPAR	PAR	NPAR
1% of data	0.956	0.998	0.970	0.986
5% of data	0.838	0.866	0.919	0.946
10% of data	0.748	0.808	0.954	0.880

Better use of PTM information

In terms of pathways, PTM status of some well-studied proteins is starting to become available, however the information is used primarily to differentiate the modified proteins from their unmodified counterparts. At the current stage this information is not sufficient to map experimentally observed changes for example in phosphorylation patterns to biological functions and regulatory activity, since phosphorylation of proteins is not entirely an on/off switch, rather what can be seen as a continuous gradient of occupancy. Furthermore, it has been shown in the literature that phosphorylation at different locations on a protein may alter its function or interaction partners. Thus it is inadequate to merely differentiate phosphorylated and dephosphorylated versions of a protein in a pathway. The goal in that regard should be to devise a computer-readable annotation method for phosphorylation events and their implications. Only with such a knowledgebase and infrastructure in place, can the quantitative phosphoproteomics fully utilize its potential.

Quantitative phosphorylation information could be better utilized in proteomics experiments. An *in silico* digestion of known human proteome yields approximately 1.7 million theoretical tryptic peptide-protein associations, within the observable mass range. Approximately 3800 degenerate peptides are annotated with differential phospho-states, thus discriminate between different proteins. In terms of functional analysis, changes in phosphosite occupancy could be a useful clue in determining regulatory activity that does not alter protein abundance.

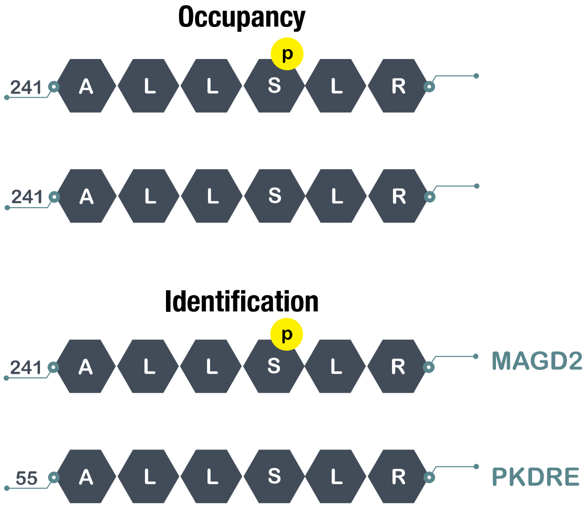


Figure 13: Site occupancy (upper panel) is an important parameter for some post-translational modifications such as phosphorylation. Changes in occupancy occur due to signaling events and is thus an alternative mechanism of regulation that does not involve changes in protein abundance. Phosphorylation status could also be interesting in handling the protein inference problem; several thousand degenerate (non-unique) peptides have differential phosphorylation evidence, which can be obtained from UniProtKB. Two different proteins (lower panel) containing the same tryptic peptide can be differentiated from each other (with a certain statistical bias) if the phosphorylated peptide has been observed in the sampe. Note that this procedure does not work the other way around; observing the non-modified peptide does not discriminate between the proteins.

Data visualization problem

Useful models of enrichment or efficient and robust computation of enrichment scores are not the only challenges in pathway analysis of expression data. Pathway analysis tools aim to identify pathways that have likely undergone regulatory activity, based on inference from expression data. However, in almost all cases the results of pathway analysis should be seen as directions for further analysis, requiring human intervention.

Given this setting, the visualization of expression data, as well as pathways constitutes a challenge that needs to be addressed. A significant amount of time and work during this thesis has been invested in pursuit of effective and intuitive visual display of data.

Visualizing pathways

Biochemical pathways are abstract concepts that are defined by various different curators and collected in many different databases, represented in various different formats such as SBML (Hucka et al. 2003) and BioPAX (Demir et al. 2010), to name a few.

Visualization of pathways is a challenging task on two different levels: to visually and intuitively display the reactions within a pathway, as well as to display the complex inherent hierarchy of pathways when attempting to display results following a pathway analysis. While the former is a well-studied problem with several solutions available such as Cytoscape (Shannon et al. 2003) or ChiBE (Babur et al. 2010), the latter problem have not been addressed adequately, to our best knowledge. Large biological networks, such as the human kinome or interactome, from complex networks that are usually referred to as a *hairball* (Lander 2010), or the somewhat whimsical term a *ridiculogram,* a term attributed to University of Michigan scientist Mark Newman. Newman defines the term as graphs that are visually stunning, scientifically worthless, and published in Science or Nature.

Pathways are a series of biochemical reactions that occur in connection with one another, fulfilling a particular function in the studied system. Larger, more complex functions such as gene expression or DNA repair can typically be divided into smaller pathways with more specific functions, and those sub-pathways can in turn be further divided, or branched into even more, giving rise to a hierarchical structure where pathways could be entirely or partly overlapping with each other with respect to the proteins they include. This hierarchical structure could be a source for useful insights when coupled with results of pathway analysis tools. In Paper I, we demonstrate this idea with visualization, developed as a Java applet, which combines the hierarchical structure of pathway data with the multi-model pathway analysis. In this visualization each unique set of proteins is a node in the graph, with directed edges connecting super-pathways to their children. Nodes are coloured with a two-dimensional colouring scheme to indicate the significance scores from respective model (Figure 14).

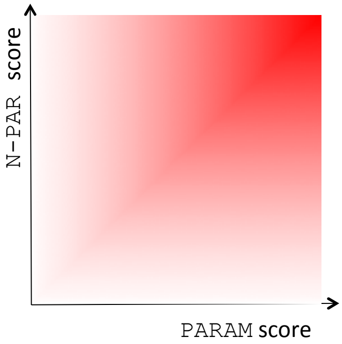


Figure 14: Representation of the colour scheme used to display multi-model scoring scheme. Use of two different models and subsequent scores make it difficult to intuitively display enrichment information on hierarchical graphs. Using linear blending among primary colors provides an intuitive display of the available information: red indicates pathways that are highly significant according to both models, yellow and magenta with a single model. Nodes that lack colour (white) are insignificant with both models.

While this approach circumvents the necessity of identical nodes (i.e. pathways that have the same set of proteins) it also implies that some nodes contain more than one pathway, and that a node might have more than one parent. The resultant graph is formally a *hypergraph* and not a standard hierarchy. Nevertheless, the visualization has been useful in identifying regulatory activity as shown in Figure 4 in Paper I. Analysis of the ductal-lobular dataset (Paper I, p. 2961) reveals highest significance on the *Transcripton* pathway, however by investigating the results using the hierarchical visualization, it is possible to identify RNA polymerase II activity as the likely target of regulation. Since every protein in a pathway exists in the parent pathway as well, a cluster of highly regulated proteins would give rise to better enrichment scores not only in that pathway but also for the parent pathways, making the hierarchical overview tool a valuable asset in pinpointing spots of functional expression regulation.

Visualizing datasets

Given a dataset consisting of several thousand proteins, what is the optimal way of displaying all that data in an intuitive way? Datasets can contain several important artefacts that might not be noticeable when displayed as a large table. Such artefacts could be an unexpectedly large number of missing values, unbalanced up/down-regulated proteins, unexpectedly many proteins with outlying expression ratios, just to name a few.

Artefacts like these might indicate a systemic bias, possibly originating from the sample preparation step. For instance in a standard comparative expression analysis, the expectation would be to have median ratio close to 1, that is approximately as many proteins would be up-regulated as down-regulated between the samples, given that equal amount of proteins were taken from each sample. If the majority of the ratios have shifted to either direction one might question whether or not there were equal amount of material in each condition. Similarly distribution of ratios and p-values in the dataset might indicate potential bias introduced by erroneous steps in prior data analysis.

Furthermore, since many proteins are active in a number of different pathways, by utilizing the protein-pathway associations it could be possible to provide visual overview of potential clusters with high, or low, expression regulation. For this purpose, we set out to develop a visualization method to provide an overview of the dataset using d3.js JavaScript library.

The first approach was to use a hexagonal grid layout with the goal to encode three key pieces of information for each protein in a visual manner; with each protein displayed as a node, the ratio displayed as colour, the ratio significance (calculated from replicates) as colour intensity and functional proximity (defined as the number of pathway associations shared) as the distance between the nodes such that proteins that commonly occur together in pathways would be closer to each other (Figure 15).

The layout algorithm for placement of the nodes on the grid was formulated as a spiral layout, placing the first node in the centre hexagon and continuing in a growing spiral. In connection with the layout we define the optimization problem where the goal is to minimize the objective function, defined as:

$$E = \sum_{\forall x,y} n_{x,y} \cdot d_{x,y}^2$$

where $n_{x,y}$ is the number of pathways that protein pair (x,y) are both a part of, and $d_{x,y}$ is the distance between the pair of nodes representing these proteins. This objective function represents some form of information entropy of the system, i.e. it constitutes a crude measure of order/disorder. The optimal solution to this problem, that is the ordering of elements that minimize *E*, should yield a layout where proteins that co-function with each other are placed very close to each other. The minimization of the objective function is not a trivial task however, given a dataset of consisting of 2000 proteins the number of different layouts, which is equivalent to the number of permutations, has more than 5000 digits and thus beyond any exhaustive search attempt.

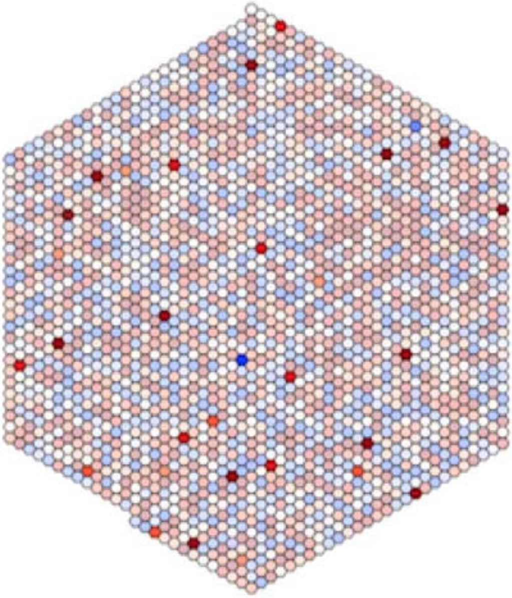


Figure 15: Hexagonal grid visualization, where each hexagon represents a protein and its color represents the ratio (red: up-regulation, blue: down-regulation). The distance between two nodes is associated with the number of pathways that particular pair of proteins are both involved in. Proteins of a chosen pathway highlighted, showing that the layout algorithm was not effective enough to bring proteins in this pathway close to each other.

Genetic algorithms (GA) are a subclass of evolutionary algorithms (EA) are a family of population-based metaheuristic methods that attempt to mimic evolution and natural selection process to generate optimization and search problems (Mitchell 1998). It has been postulated that deterministic global optimization methods are computationally too demanding for most biological modelling problems and that evolutionary strategies are better suited for this purpose (Szallasi, Stelling, and Periwal 2010).

Standard approach in GA optimization is to represent potential solutions to the problem as individuals and define the fitness function based on the objective function that is to be optimized. Starting with a random population, as the evolutionary processes such as reproduction, mutation and selection, take place the individuals strive to increase their overall fitness and thus approach the optimum solution for the objective function.

As the solutions to this particular problem are ordering of the elements in the dataset, certain aspects of standard GA approach are not feasible. For instance, bisexual reproduction of individual solutions, where new solutions inherit half their "genome" from one parent and the other half from the other, is not possible since each element needs to appear once and only once. Without bisexual reproduction, or multiple inheritances, randomized crossover events are also not implementable in this case.

However a significant portion of the biomass on our planet consist of bacteria which have evolved into many different niches and adapted to extreme conditions, without the diversity provided by bisexual reproduction. We have thus adopted a GA where we model bacterial evolution; the genome of each individual being a permutation of proteins in the dataset and the asexual reproduction process heavily influenced by various mutation events. In order to provide as much variety as possible several different types of mutation events were modelled, mimicking processes like point mutations involving reshuffling portions of the genome, positional swapping of portions of the genome resembling transposons in eukaryotes. Furthermore, for further variability and avoiding local optima, each generation had "newcomers" introduced, based entirely new and random permutations of the initial dataset.

Given the high mutational hazard on individuals, stability of best solutions was a problem. This problem can however be addressed rather simply by addition of elitism heuristics, where the most fit individuals each generation are immune from high level of mutations since they are deemed to be better adapted to their environment, whereas less fit individuals are under higher selective stress.

While promising initially the GA-based optimization soon showed to be unpractical for the intended purpose, as the optimization took longer time to run than the actual enrichment analysis process. The results that could be achieved within acceptable number of generations, or time duration, were unfortunately not satisfactory (see Figure 15). Even though further tweaking of the algorithm could potentially increase the speed of convergence, for the project was scrapped in favour of other projects due to time constraints. Pathway oriented optimal layout of datasets is not an easy task, especially at runtime, without prior optimization. Therefore later efforts in dataset optimization have shifted from functional grouping of the proteins, towards more overview type visualizations.

Figure 16 depicts the dataset visualization that has been implemented in newer versions of FEvER tool. This visualization is based on a chart type called bubble charts, where each entity is displayed as a bubble, its colour intensity representing the expression ratio and its size representing the significance. The chart provides a simple and quick overview to finding most interesting data points (i.e. large bubbles with bright colours), as well as displaying the proportional representation of missing values, which are displayed as small, fixed-size, black bubbles.

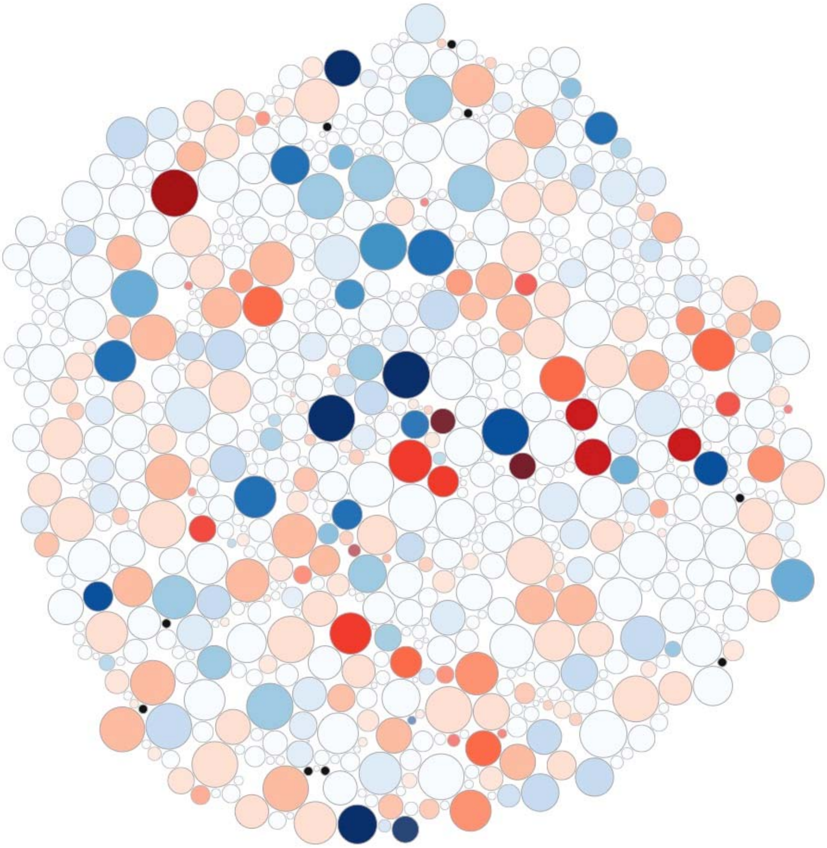


Figure 16: Bubble charts visualize datasets by using bubbles, or circles. Size and colour of the circles represent the significance and magnitude of the expression ratios. Location of the circles, however, is not of any particular meaning.

Cancer & Cancer Proteomics

> "Tumors destroy man in a unique and appalling way, as flesh of his own flesh which has somehow been rendered proliferative, rampant, predatory and ungovernable. They are the most concrete and formidable of human maladies, yet despite more than 70 years of experimental study they remain the least understood."
>
> Francis Peyton Rous, Nobel lecture, 1966

There is no doubt that cancer has become the biggest medical burden globally, surpassing coronary heart disease or stroke in number of deaths caused, with the trends indicating increasing numbers for both new incidents and mortality, over the coming years (Ferlay et al. 2015). In Sweden, the official estimates state that approximately every third individual alive today will be diagnosed with cancer at some point during their lifetime (Bergman, Hont, and Johansson 2013). The numbers in the United States are similarly alarming, where the lifetime probability of being diagnosed with invasive cancer is estimated to be around 40% (Siegel, Miller, and Jemal 2015).

Given that the "war on cancer" was initiated over 40 years ago epidemiological studies point towards no significant victory towards this adversary, despite the remarkable progress made in increasing our understanding the molecular mechanisms underlying the pathogenesis and in some cases disease progression (Hanahan 2014).The critical question is then why do we keeping losing ground in this war, despite the extraordinary resources that have been utilized for this purpose. The enemy is spreading to new areas around the globe, growing in numbers. So one might wonder the reasons behind how the cure for cancer eluded us so well.

One possible answer to that question is that the prevalence of cancer is associated with increasing global population together with growing life expectancy, as well as wider adoption of cancer-associated lifestyle choices, such as tobacco and alcohol consumption, obesity, physical inactivity, and increasing share of processed foods, sometimes referred to as "western" diets (Jemal et al. 2011; Ferlay et al. 2015).

Another important insight is that the enemy might not be one single adversary but rather a collective term by which many different conditions described. Malignant tumours present in a wide range of forms that may arise from various different cell types around/in the tumour resulting in many different pathological states (Weinberg 2013). That being said, most cancer types, if not all, share a number of characteristics, from their origin to progression, that allow them to survive and adapt to changing conditions. Understanding these common mechanisms is likely to be vital in our ongoing war against cancer.

The Hallmarks of Cancer

In their seminal review, *The Hallmarks of Cancer*, Hanahan and Weinberg point out that most, if not all, types of human cancer exhibit a number of common traits, as a collection of capabilities acquired throughout the multi-step progression from a normal state to malignancy (Hanahan and Weinberg 2000) typically involving the loss of function or gain of function mutations associated with critical genes. They postulate that while exact order in which these cells acquire these traits may vary, the acquisition of six main traits collectively orchestrates the tumorigenic transformation. These being: development of self-sufficiency in growth signalling, sustained replicative immortality, insensitivity to growth-inhibitory signalling, evasion of programmed cell death, potential for inducing angiogenesis as well as tissue invasion and the ability to metastasize (Figure 17).

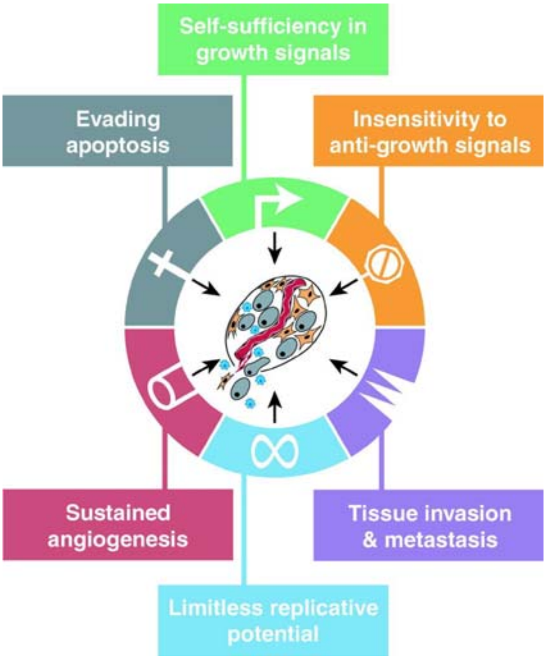

Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

Figure 17: Original hallmarks of cancer by Hanahan and Weinberg (Hanahan and Weinberg 2000), reproduced with permission from the publisher.

A balance between growth inducing and inhibiting factors, and the signalling events that govern survival and apoptosis regulates the lifecycles of normal cells. Tumorigenic cells liberate themselves from dependence on exogenous signalling for growth and survival, as well as avoiding mechanisms that inhibit cell growth, replication and trigger apoptosis. Together these attributes allow cells to grow and expand in numbers. However, without supporting processes in place for such a growth-driven mechanism cells cannot sustain limitless expansion, as the replicative potential of cells is intrinsically limited in normal conditions. Telomere maintenance mechanisms, such as up-regulation of telomerase expression, provide cells unlimited replicative potential. Another supporting process that needs to be established for the tumours to grow is the sustained angiogenesis. As cells grow and replicate the access to vital resources decrease while the demand for these resources increase dramatically. Thus for sustained growth cells need to promote growth of new blood vessels. Finally, as the nutrients and space become limiting factors, tumours invade nearby and distant tissues.

Hanahan and Weinberg later amended their initial proposition adding two more hallmarks as well as two "enabling characteristics" (Hanahan and Weinberg 2011). One of these enabling characteristics is genomic instability. Cancer as a disease has its roots in genomic deregulation and it is no surprise that neoplastic cells exhibit some genomic defects. However, tumour cells typically have defects in the complex DNA maintenance machinery, thus increasing the rate by which mutations occur. Mutant copies are passed on either due to these defects, or due to insensitivity to apoptotic signalling events, another hallmark trait of these cells.

Two of the remaining three characteristics (Figure 18) are related to the complicated relationship between cancer and the immune system. It has been shown that inflammation may have tumour-promoting effects, helping the tumour secure growth factors and survival signalling, as well as modifying the extracellular matrix that may facilitate angiogenesis, invasion and metastasis. Meanwhile, the immune system is also responsible for the continuous surveillance that leads to hindrance or eradication of tumour formation and progression, as well as micro-metastases. This effect of the immune system has been shown both in animal models and clinical epidemiological studies of human cancers. Neoplastic cells thus have an intricate relationship with cells of the immune system, where they benefit from the tumour-promoting inflammatory processes, which may facilitate their acquisition of more hallmark traits, and simultaneously try to evade the selective destruction in connection with the immunological surveillance.

Lastly, cancer cells typically exhibit altered energy metabolism compared to their normal counterparts. This often occurs in the form of cells resorting to the much less efficient glycolysis instead of oxidative phosphorylation, an anomaly called *the Warburg effect*. While the functional rationale behind this effect is poorly understood, one theory is that glycolysis provides advantages in driving large-scale biosynthetic processes that are fundamental to cell growth and replication. Furthermore, hypoxic conditions that are prevalent in most tumour environments could also contribute to the equation, where subpopulations of cells with differing energy metabolisms may have symbiotic relationships. Nevertheless, reprogramming of the energy metabolism appears to be a common trait of many cancers.

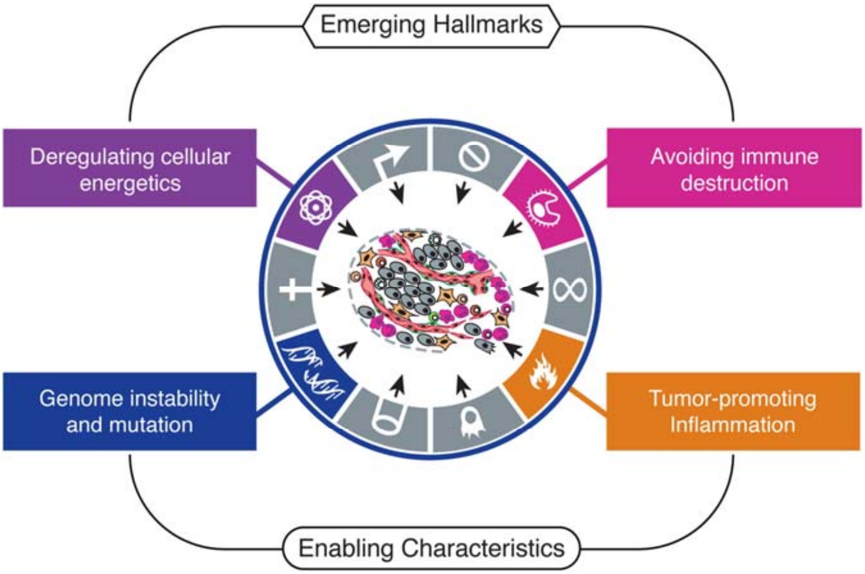

Emerging Hallmarks
Deregulating cellular energetics
Avoiding immune destruction
Genome instability and mutation
Tumor-promoting Inflammation
Enabling Characteristics

Figure 18: Emerging hallmarks and enabling characteristics from the revised version of the Hallmarks of Cancer by Hanahan and Weinberg (Hanahan and Weinberg 2011), reproduced with permission from the publisher.

An important advantage of highlighting these common characteristics, or hallmark capabilities, of tumours is that it provides targets on which to concentrate the development of therapeutic efforts. Referring to the war analogy, if the mechanisms by which the enemy sustains its function can be identified and targeted, the battleground may look different in the future. These hallmarks capabilities provide that type of understanding and potential targets, by which therapeutics can eliminate or mitigate vital functional aspects of neoplastic cells. Examples of such therapeutics include telomerase inhibitors to limit the replicative potential of these cells, inhibitors of VEGF or EGFR signalling to prevent angiogenesis and proliferation (See Figure 6 in (Hanahan and Weinberg 2011)). However, given the heterogeneity of tumours the selective pressure resulting from treatments directed at specific targets is likely to result in relapse either by means of adapting to the treatment by decreasing dependence on the treatment target or by expansion of a minority subpopulation which is insensitive to the treatment.

Expression profiling: the biomarker question

The term *biomarker*, or *molecular marker*, is loosely used for molecules that are measurable indicators of a specific biological process, condition or disease. Proteins have a unique position for biomarker discovery, as they are the ultimate effectors of biological processes in connection with disease, response to treatment and recovery (Rifai, Gillette, and Carr 2006). In theory, biomarkers are valuable tools for three main purposes; for gaining insights for accurate and early detection of disease (*diagnostic* biomarkers), for prediction of disease progression, recurrence and survival (*prognostic* biomarkers), and for predicting response to different therapeutics (*predictive* or *response* biomarkers) (Frantzi, Bhat, and Latosinska 2014). All three types of biomarkers are vital in the war against cancer. Early detection and accurate classification of the disease is strongly correlated with longer survival, indicating the importance of diagnostic biomarkers, while identification of patients under higher risk and optimal course of therapy of a given individual provides the motivation for searching prognostic and predictive markers, respectively.

However despite the abundance of studies aimed at identifying biomarkers in different cancer subtypes, the clinical utility of the candidate markers have not in large realised the intended goals (Hanash 2011). In fact, recent estimates show that less than 1% of reported cancer biomarkers are actually useful in clinical practice (Kern 2012). While the exact reasons for this lack of efficiency may be difficult to pinpoint, the biomarker problem has been extensively discussed and reviewed in literature (Rifai, Gillette, and Carr 2006; Sawyers 2008; Kern 2012; Frantzi, Bhat, and Latosinska 2014; Kondo 2014). In short, three main types of challenges impede development and use of biomarkers; these are challenges in discovery candidate markers and the validation of these candidates.

In order to be practically useful, biomarkers should preferably be measurable reliably and efficiently (in terms of time and cost), as well as through non-invasive means. This definition immediately poses a challenge in cancer settings, as it is difficult to reach tumours in a non-invasive manner (Sawyers 2008). While biofluids such as plasma provide non-invasive means for discovery efforts, the complexity and dynamic range of sample proteome impedes development of reliable methods of discovery. In connection with the complexity and dynamic range, many likely candidates of disease specific markers are expected to have low-abundance in the samples, body fluids and tissue samples alike. Furthermore, in many cases there is significant individual variation, both on human and disease level.

Once discovered the validation of candidate biomarkers poses a challenge in itself. With the technological and methodological advances, it is now possible to characterize, at an unprecedented level, the changes in DNA, or gene expression level, which provided important insights to underlying mutations and/or genomic defects in particular conditions. However it is now well established that gene expression and protein expression do not necessarily correlate, as demonstrated in Paper IV, supporting the lack of transition to clinical use for many of these markers. The difficulty in working with solid tumours is another hinder to be tackled, specifically for predictive markers, since it is often impractical to take several biopsies from patients who are already under considerable burden (Sawyers 2008).

An important aspect of biomarker discovery studies is the problem formulation and subsequent data analysis. A common approach in discovery-based gene or protein expression studies is to look for entities that have statistically significant differential expression, and then reporting these entities as potential biomarkers in discriminating between the conditions studied. This can be considered as a supervision bias or a self-fulfilling prophecy. Similar potential pitfalls of fallacies in biomarker discovery have been voiced by Kondo and Kern (Kondo 2014; Kern 2012), and several strategies in validation of candidate markers are proposed (Rifai, Gillette, and Carr 2006; Frantzi, Bhat, and Latosinska 2014).

Finally, virtually all biomarker studies are carried out in isolation, that is to say they focus on a single disease or tissue, and ignore the possible interference from other organs producing these biomolecules. Many proposed biomarkers appear in very diverse disease states. This is merely a reflection of their roles in dealing with stress in a cell and hence many lack any specificity beyond the recognition that there is generically something wrong with the patient. Furthermore, the heterogeneity of expression between patients is problematic, there appears to be no common baseline level that can be applied. This restricts many markers to a monitoring role, assuming one can have a reading from the patient when they are disease free. A more integrative approach is needed with many more "healthy" controls.

Despite falling short of expectations, biomarker discovery will likely stay as one of the most common motivations for shotgun proteomics experiments.

Cancer datasets in focus

In this thesis we present four studies (Papers II-V), which were aimed at investigating the proteomes of several different types of human malignancies using mass spectrometry-based shotgun approaches, highlighting potentially key pathways in pathogenesis and progression with bioinformatics tools presented in the previous chapter.

Soft-tissue sarcoma

Sarcomas are highly malignant and heterogeneous tumours of mesenchymal origin. Despite representing less than 1% of human malignancies (Baird et al. 2005), there are more than 100 subtypes of soft-tissue sarcomas (STS) recognized by histological examination (van de Rijn and Fletcher 2006). Approximately one third of the patients develop metastases, and metastatic disease at the time of diagnosis is not uncommon (Ferguson et al. 2011).

Poorly differentiated sarcomas represent a diagnostic challenge as these tumours lack a typical and easily identifiable phenotype. A particular category of poorly differentiated tumours is pleomorphic sarcomas, which in turn have multiple subtypes including the malignant fibrous histiocytoma/undifferentiated pleomorphic sarcoma (MFH/UPS). While there are several studies regarding the genomic profiling of this disease, little is known about the proteome. In paper II, we describe a study where we investigate the proteomes of STS samples collected at Lund University Hospital. We have focused in particular on leiomyosarcomas (LMS) and MFH/UPS examining the possibility that these two subtypes share a common lineage.

Starting out with an analysis of various types of STS (139 tumours in total) based on 2D-DIGE experiments, we could not find any meaningful results using unsupervised hierarchical clustering. Considering only the LMS and MFH/UPS samples, some distinct clusters of LMS tumours did emerge, however the MFH/UPS samples were could not be distinguished from the LMS samples. We identified 20 tumours, 15 of which representing the three clusters of LMS and a group of 5 MFH/UPS samples, which were analysed by both tandem-mass tagging (TMT) and label-free LC-MS/MS. In a multi-group comparison we identified vinculin (VINC), collagen type VI alpha-3 chain (COL6A3) and myosin heavy chain smooth muscle isoform (MYH11) as discriminators between the 4 groups. Both VINC and COL6A3 have been associated with cancer related pathways in different malignancies in the literature, particularly of interest both have previously been reported to undergo expression regulation in mesenchymal tumour cells (Schreier et al. 1988; Schenker and Trüeb 1998).

In depth analyses based on data from label-free LC-MS/MS experiments revealed 2 distinct subgroups amongst the LMS samples in the study (Paper II, Figure 3 and Table 3). One of these groups is significantly enriched in ribosomal proteins as well as subunits of eukaryotic translation initiation factors. Survival data was not available for these samples, thus it was not possible to carry out survival analysis to investigate whether or not the subtypes differ in malignancy. Functional analysis of the 156 discriminating proteins revealed Granzyme-A signalling in apoptosis and survival, cytoskeleton remodelling and telomere maintenance as the most targets of expression regulation.

Despite the emergence of LMS subgroups with respect to protein expression, MFH/UPS samples could not be discriminated as a group from the LMS samples. Instead they were spread between the two clusters (Paper II, Figure 2). It has been shown that human mesenchymal stem cells (hMSCs) can give rise to MFH/UPS samples via commitment to *Wnt* signalling (Matushansky et al. 2007), and that both subtypes of sarcomas can originate from the same murine model (Guijarro et al. 2013) via *Notch* signalling. In the light of these studies, we believe our result indicate that tumours classified as MFH/UPS most likely arise from hMSCs during different stages of smooth muscle differentiation. This discovery-phase study opens the door for further functional analyses of these rare and aggressive tumours.

Gastroesophageal tumours

Oesophageal tumours are rare and malignant tumours that arise in the epithelial lining of the oesophagus. While the overall occurrence is about 1% of all human malignancies, the disease is about four times more likely among men than women (Siegel, Miller, and Jemal 2015; Bergman, Hont, and Johansson 2013). With the 5-year survival rates reported at just below 20%, oesophageal tumours are among the top 5 causes of cancer related deaths in males between ages 40-59 in the U.S. (Siegel, Miller, and Jemal 2015). In Sweden, the number of incidents has been on a slow but steady rise over the past 30 years, and the 5-year survival rates are about 13%, for both sexes (Bergman, Hont, and Johansson 2013). While the standard treatment has been surgery, optimal course of action in treating these tumours is still a matter of debate (Mariette, Piessen, and Triboulet 2007).

In Paper III, we report a study that aims at highlighting mechanisms behind malignancy as well as potential biomarkers for predicting prognosis and response to alternative treatment methods, a need expressed previously in literature (Tew, Kelsen, and Ilson 2005; Koshy et al. 2004). We profiled the proteomes of 81 gastroesophageal tumours, paired with control samples taken from healthy tissue from the same patients (see Paper III, Table 1), using both 2D-DIGE and shotgun mass spectrometry.

Based on protein expression patterns, we observed a clear separation between tumours and normal samples, using unsupervised clustering methods. Proteins discriminating the samples ($q < 0.01$) were enriched for the GO biological processes such as; epithelial cell differentiation, ectoderm development, epithelium development, keratinocyte differentiation, epidermis development, epidermal cell differentiation, peptide cross-linking and cellular homeostasis.

Furthermore, we carried out pathway and GO-enrichment analysis using FEvER method (Paper I), comparing the tumours to normal samples. We filtered the dataset to the proteins for which a tumour/normal ratio can be calculated for at least 3 individuals. We could thus do a non-parametric, paired test of significance for the observed ratios. Based on this filtered dataset, we identified several extra cellular matrix related pathways as the most likely targets of expression regulation. Specifically we found that Periostin precursor (POSTN) and Cytokeratin-7 (KRT7) were upregulated 169x (adjusted p-value = 0.00022) and 111x (adjusted p-value = 0.01838) respectively, points towards significant restructuring of the cytoskeleton and cell-adhesion interactions. Among other high scoring pathways we observed *Translation, Post-*

translational protein modification, Syndecan interactions, ECM proteolycans, Gene expression, Collagen formation, and *Asparagine N-linked glycosylation*.

One of the aims of this study was to investigate the oesophageal tumour proteome for prognostic markers of the disease. We carried out survival analysis based on the clinical parameters as well as the protein expression data, however we could not identify any proteins that were indicative of higher risk or more aggressive disease. Specifically, there were no proteins in our dataset that could discriminate the patients with the longer survival times from those that were shorter.

Breast cancers

Breast cancer is one of the most common malignancies and one of the leading causes of death among women and as such it is one of the most studied forms of cancer. Globally, breast cancer ranks second in prevalence, just behind of lung cancer, and fifth in mortality (Ferlay et al. 2015). In Sweden, breast cancer constitutes over 30% of the diagnosed malignancies for women, however with developments in treatment and earlier diagnosis, the 10-year survival is approximately 80% (Bergman, Hont, and Johansson 2013).

This thesis contains two studies aimed at obtaining a better understanding of the breast cancer proteome and functional regulatory activities, both in terms of pre-clinical research, and potential therapeutic application developments in mice models. Despite the extensive efforts in profiling the genome and transcriptome of the disease, few molecular biomarkers have made it to clinical use, and stratification typically depends on histological parameters like the TNM (Tumor Node Metastasis) status. The commonly used molecular markers are the expression of receptors of estrogen (ER), progesterone (PR), and the receptor kinase ERBB2 (also known as HER2/*neu*). Several panels of molecular markers, based on transcriptomics studies, have been proposed for classification of breast cancer in several subtypes (Perou et al. 2000; Sørlie et al. 2001; Sorlie et al. 2003; Hu et al. 2006).

In paper IV, we describe a study aimed at investigating the similarity of human breast cancer cell lines (BCCLs), a commonly used model system, to the corresponding *in vivo* human molecular subtypes that they are assumed to represent. The study revealed several interesting insights; one of which is the observed similarity of the BCCL proteomes. In fact, just about two thirds of the identified proteins (2270 out of 3417) were found in all five cell lines that were investigated, and an additional 521 proteins (approximately 15%) were shared among four out of five cell lines (Paper IV, Figure 1). This finding is supportive of the idea that cell lines predominantly exhibit a generally similar "core" proteome that allows them to survive and grow under *in vitro* conditions.

Furthermore, comparisons of proteome and transcriptome of the same cell lines provide significant differences. Not only do our results confirm the notion of a general lack of correlation between abundances of mRNA and the corresponding proteins, but also the cell lines cluster differently based on expression data on different levels (Paper IV, Figure 4). Functional analyses using the FEvER method, described earlier, and in Paper I, we could investigate the similarities, or lack thereof, between the pathways that are likely to be differentially regulated in these cell lines. Again, an overall lack of correlation was observed between the significance scores for the pathways between proteomics and transcriptomics datasets.

Lastly, we compared the proteomes of 450 tumour samples to the five BCCLs in the study. While a comparable amount of proteins identified in both cases, an intriguingly small overlap was observed between the datasets. Suspecting that the disparity might be caused by the contribution of tumour stromal cells, the addition of adipocytes and a cell line derived from human breast fibroblasts was devised. However, while increasing the number of identified proteins by 300 new proteins, the addition of these cells did not add to the overlap of proteins between cell lines, and the clinical samples. GO term analysis of the datasets highlighted the difference of processes that are activated in the different conditions; the tumour samples had much better coverage of cell adhesion and hormone signalling pathways while the cultured cells had a similarly better coverage of metabolic processes and cell-cycle regulation. Tumour samples are very heterogeneous and highly complex, containing cells of different lineages, such as blood cells or cells of the immune system (Weinberg 2013). In comparison, cultured cell lines are very homogeneous, which is a direct consequence of their clonal expansion. Furthermore, cells in a culture not only share essentially the same genome, but also the same environmental factors originating from culturing conditions and are not exposed to the same selective pressure dynamics that is present *in vivo*; they simply grow and divide. Considering these results we conclude that while being a useful model of understanding the basic cellular processes in different types of cells, it is unlikely to be a suitable model for biomarker discovery or therapeutic models.

From a more clinical point of view, hormonal involvement in breast cancer has been reported as early as the end of 19th century (Beatson 1896). Hormone receptor status, that is whether or not malignant cells express oestrogen and progesterone receptors, is an important parameter in the therapeutic decision-making process. In particular ER+ tumours, among the most common (Patani, Martin, and Dowsett 2013) and diverse clinical groups (Cancer Genome Atlas Network 2012), often rely on oestrogen signalling for growth and are susceptible to hormonal therapy. In Paper V, we describe a study with which we target the dependence on oestrogen receptor signalling in ER+ breast cancer, using a patient-derived, luminal-like mouse xenograft model (Bergamaschi et al. 2009; Skrbo et al. 2014). Our results showed a clear dependence of the malignant cells on oestrogen signalling, more than 50% decrease in tumour volume was observed over 2 weeks of treatment, either by oestrogen withdrawal or by a ER-antagonist pharmaceutical, *fullvestrant* (see Paper V, Figure 2a). The estradiol concentration in serum was measured for control of the hypothesis that fullvestrant treatment would block ER-signalling even in the abundance of the agonist (Paper V, Figure 2b). Furthermore, we analysed the proteomes of tumours that have undergone both types of treatments, as well as control, and carried out pairwise comparative pathway analysis using the FEvER tool (Paper I) and found signs of metabolic regulation. Treatments were compared to control separately, and combined, yielding three pairwise comparisons. In all three cases, *TCA Cycle and respiratory electron transport* and *metabolism of lipids and lipoproteins*, especially *fatty acid, triacylglycerol, and ketone body metabolism,* pathways were found to be highly significant (Paper V, Table 1). In both cases, the majority of the proteins in these pathways were found to be upregulated in comparison to untreated tumours, and together these results indicate a likely increase of oxidative phosphorylation activity upon oestrogen-signalling inhibition (Paper V, Figure 3a). These results were confirmed by both GO annotation enrichment applied to the same dataset (Paper V, Figure 3b) and by high-resolution magic angle spinning magnetic resonance spectroscopy (HR MAS MRS), an orthogonal method of metabolite quantification (Moestue et al. 2011), on a new set of PDX samples (Paper V, Figure 4). Mass spectrometry proteomics, together with the in depth metabolomics is strongly indicative of a therapy-induced decrease in the previously discussed *Warburg effect*, and a reprogramming of cellular energy metabolism upon oestrogen deprivation.

Finally, investigation of subpopulations previously defined in the same PDX model (Skrbo et al. 2014) revealed that the oestrogen-treatment has an effect on proportional representation of these subpopulations in tumours (Paper V, Figure 5). Furthermore, proteomic analysis of these subpopulations indicate that the metabolic differences are likely to exist in between two of the subpopulations, and that the treatment changes the subpopulations by means of selective pressure. Taken together, our results from Paper V give insights regarding *in vivo* molecular and functional dynamics as a result of hormonal therapy.

Outlook

Where do we stand today?

Despite the hiccups, e.g. (Ransohoff 2005), in the early years of the century, proteomics as a field has established itself to the extent that the benefits of studying the proteome is widely accepted and recognized. While one cannot consider proteomics is a fully matured field of biomedical science, the remarkable progress made in the past 15 years should be recognized.

Currently available technologies together with well-designed workflows have provided a deeper and more consistent view of the human proteome than ever before. We can now identify and quantify several thousand proteins in discovery-phase shotgun experiments, efficiently and reproducibly quantify over a hundred proteins in targeted studies, and identify thousands of post-translationally modified peptides from a single sample, in a single experiment. On a bigger scale, drafts of the human proteome have been assembled and published last year, by two independent research groups (M.-S. Kim et al. 2014; Wilhelm et al. 2014), which constitutes a significant step towards a more thorough understanding of how our bodies function. Furthermore, not only do we know which proteins are expressed in our body, but also where they are expressed, as a study on tissue-specificity of thousands of proteins described recently (Fagerberg et al. 2014). These landmark achievements in proteome research, will no doubt prove to be invaluable in expanding our understanding of complex diseases like cancer, paving the way for advance therapeutics, tailored for the needs of each individual patient.

As discussed in the previous chapter, cancer is not one single disease, thus a universal cure for cancer is likely to remain elusive. The idea of considering each case unique has gained momentum. Along this line of thinking, there is no one cure for cancer but rather the goal is to find the right therapy for the right patient. A concept, usually referred to as *personalized medicine*, implies characterization of critical mutations, identification of the amplified or silenced pathways and other regulatory activity, for each and every patient.

Over the past decade, remarkable progress has been made toward personalized medicine, as patients are screened for molecular biomarkers in many types of cancer. In breast cancer specifically, HER2 status has transformed from a prognostic biomarker to a predictive biomarker. Introduction of *Herceptin* (Trastuzumab) for HER2$^+$ breast cancer patients, *Gleevec* (Imatinib) for BCR-Abl$^+$ chronic myelogenous leukemia

(CML) patients as well as PARP inhibitors for BRCA1/BRCA2 mutant breast or ovarian cancer patients, show great promise in management of these malignancies.

There are still limitations, to the current methods, and obstacles that need to be overcome. These shortcomings are of both technological and biological nature, for example mass spectrometry proteomics is still inadequate in sampling proteins in the lower-end of the dynamic range, especially in shotgun experiments. This particular issue constitutes an undeniable obstacle in the way for biomarker discovery studies, as discussed earlier in the Proteomics chapter. Emergence of new approaches that aim to combine the advantages of affinity-based methods and mass spectrometry proteomics (N. L. Anderson et al. 2004; Olsson et al. 2012; Säll et al. 2014) are likely to mitigate the problem to some extent.

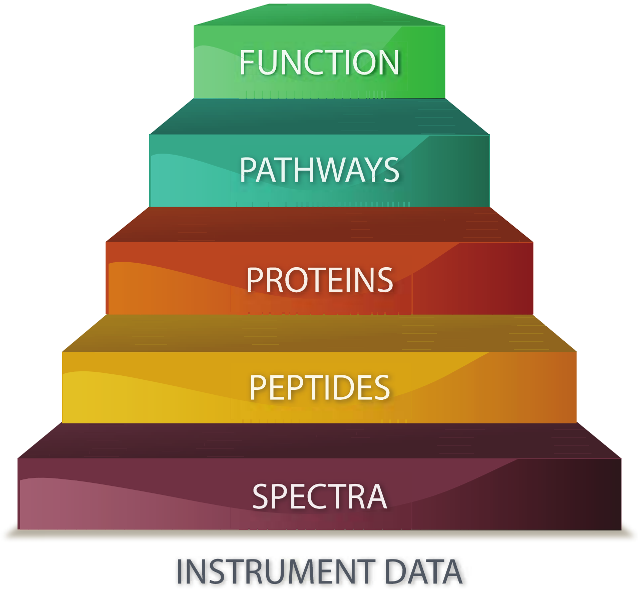


Figure 19: Steps of inference; from ground truth that is instrument data to functional interpretation and new biological insights. Each step ”up” implies inferential uncertainties and reliance on external data resources.

Furthermore, proteomics workflows heavily rely on canonical information stored on databases, whether that is on genes, peptides, proteins or functional annotations, such as pathways. This reliance on canonical data implies two potential pitfalls: the quality as well as the relevance, or usefulness, of the information available. Incorrect or misleading information on these pathways propagate due to this heavy reliance, which may prove to be detrimental when drawing conclusions on functional level, after several layers of inference.

Concluding remarks

The research covered in this thesis aims to better utilize quantitative proteomics data from mass spectrometry experiments, in order to improve our understanding of functional implications of expression regulation in cancer. The introductory chapters of the thesis provide an overview of the technological, computational and biomedical aspects of the work presented, respectively.

Paper I describes a novel method for evaluation expression regulation on a functional level, and highlight pathways that are likely to be affected by changes in protein regulation. This method is based on a relatively simple mathematical model, which in turn is built on observations originating from the relationship between physical entities such as proteins and artificial constructs such as pathways.

Papers II-V are focused on biological questions regarding the cancer proteome and protein expression regulation in different malignancies. Paper II presents an investigation of soft-tissue sarcoma proteome, specifically leiomyosarcomas. Despite poor overlap between datasets identified by different methods, we manage to highlight several interesting proteins and pathways that may likely give insights to origins and differentiation patterns amongst these rare and highly malignant tumours. A similar study is presented in Paper III, in which we investigate protein expression profiles of gastroesophageal tumours. Using the paired clinical samples, we investigate the differences in protein expression between normal esophagus and tumour samples, as well as highlighting the expression regulation for proteins in the cytoskeleton modelling and cell-adhesion pathways.

Papers IV and V are focused on breast cancer models. In Paper IV we present a study investigating the level of similarity between immortalized, patient-derived breast cancer cell lines and the tumour types these cell lines are assumed to represent. The low level of similarity between the cell lines and tumours, as well as the differences between protein and gene expression profiles raise a warning flag for inference on cancer proteome based on results from *in vitro* experiments. Our results indicate that while immortalized cell lines constitute a useful model they are not entirely representative of the molecular dynamics inside a tumour, and thus may not be a suitable model for studies for biomarker discovery.

In Paper V we describe a study, which aims to investigate dependence to oestrogen signalling in ER+ tumours. Based on the mouse PDX model we described the positive effects of blockage of this pathway, highlight the changes in protein expression and cellular subpopulations caused by the treatment, and identify changes to the energy metabolism and validate the results by an orthogonal method measuring the metabolites in the TCA cycle. Based on these results we can speculate about the dynamics between ER signalling, proliferation and energy metabolism, which opens the doors for combined therapies that are designed to hit multiple pathways, for ER+ breast cancer patients.

Taken together, the works presented in this thesis show the potential of mass spectrometry proteomics, coupled with well-designed data analysis practices. Mass spectrometry proteomics have typically been a technology-driven field; however as the instrumentation techniques mature the focus will inevitably shift to data analysis methods to climb what Aebersold referred to as the “mount bioinformatics” (Aebersold 2009). Given the challenges that are visible ahead, and presumably many more that are not yet visible, it is likely going to be a difficult climb up. However, considering the promise of the riches that lie beyond, the view from the summit should be worth the struggles of the metaphorical climb up.

Populärvetenskaplig sammanfattning

Den mänskliga kroppen är ett komplext maskineri med en rad olika reglersystem och kontrollmekanismer. Flera grenar av modern biomedicinsk forskning riktar sig mot olika nivåer av detta komplexa maskineri. En av dessa grenar heter *genomik*, som är läran om hur den genetiska koden som finns i varenda cell av en organism används och regleras. Slutförandet av Human Genome Project (HGP) gav det vetenskapliga samfundet en blåkopia av de molekylära byggstenarna i våra kroppar. Det överraskande resultatet är att nästan hela vår arvsmassa, det vill säga de proteinkodande regionerna i vårt DNA, delas inte bara bland alla levande människor, men också bland en överväldigande del av alla däggdjur.

Man kan då undra vilka mekanismer som står för de synliga och osynliga skillnaderna mellan oss människor. Svaret ligger delvis bland de biologiska molekyler som ligger på en högre nivå, dvs. proteinerna. Faktumet att en viss gen finns hos två människor betyder inte att de uttrycker samma protein, eller att det proteinet ifråga är funktionellt i samma utsträckning för dessa två individer. *Proteomik* är ännu en, något yngre, gren av biomedicinsk forskning som fokuserar på kvalitativ och kvantitativ kartläggning av proteomet. Proteomet definieras som den mängd proteiner som uttrycks av en cell, vävnad eller organism, vid en viss tidpunkt under specifika förhållanden. Proteomik studerar identiteten, kvantiteten eller funktionen av de proteinerna som finns i det undersökta systemet. Till skillnad från genomet, som är mer eller mindre identiskt hos de allra flesta cellerna i kroppen och bevaras så stabilt som möjligt, är proteomet otroligt dynamiskt med avseende på både tid och plats, och regleras konstant.

För att illustrera detta, betrakta en nervcell i ditt öga och en epitel cell på din hud. De ser ut och fungerar helt annorlunda eftersom de uttrycker olika proteiner, men de har samma uppsättning gener i cellkärnan. På samma sätt kan olika proteiner hittas i en cell som är i tillväxtfas jämfört med en cell som är mitt i celldelningsprocessen. Ett annat sådant exempel är förändringen i proteinuttryck för celler som exponeras för miljöfaktorer såsom starka variationer i tillgänglighet av näringsämnen eller olika typer av stress såsom syrebrist eller joniserande strålning.

Proteomikens roll i cancerforskningen

Det övergripande målet för många proteomik projekt är att analysera förändringarna i proteinuttryck mellan två eller flera tillstånd, särskilt i sjukdomar såsom cancer där storskaliga förändringar i proteomet inträffar. En djupare förståelse av dessa förändringar kommer i slutändan att hjälpa forskare att utveckla bättre läkemedel och läkare att skräddarsy effektivare behandlingar för patienterna.

Majoriteten av läkemedel som används idag mot cancer består av giftiga molekyler som påverkar främst delande celler. Målet med terapin bygger till stor del på antagandet att cancerceller genomgår celldelning snabbare och oftare än "normala" celler i sin omgivning. Cytotoxiska läkemedel orsakar allvarliga biverkningar för patienten, vilket har en stor inverkan på livskvaliteten. Tyvärr har patienten i många fall väldigt liten nytta av behandlingen, om ens någon alls, på grund av förvärvad resistens. Samtidigt, är den ekonomiska bördan av dessa behandlingar för samhället inte försumbar. Många patienter har nedsatt immunförsvar och är mottagliga för opportunistiska sjukdomar.

Personlig medicin

Med begreppet *personlig medicin* menas att skräddarsy behandlingen baserad på den specifika typ av sjukdom patienten har. På så sätt får patienten den behandling som har störst chans att ge hälsofördelar, samtidigt som de cytotoxiska läkemedlen används minimalt och patienten förhoppningsvis upplever en mindre inverkan på livskvalitén. Men för att personlig medicin ska bli verklighet måste vissa utmaningar tacklas. Framförallt, för att skräddarsy behandlingen till en patient, bör de mekanismer som påverkas av sjukdomen upptäckas. Pro-onkogena mutationer, som är de förändringarna i arvsmassan som främjar tumörutveckling eller överlevnad, måste identifieras. Nätverk av samverkande proteiner, som kallas för signalvägar, eller *pathways*, som har ändrats i de maligna cellerna, måste analyseras i detalj.

En rad viktiga framsteg mot användningen av personlig cancerbehandling i kliniken har gjorts. Idag lämnar nästan varje cancerpatient i Sverige prover som undersöks med state-of-the-art instrument, som avslöjar många viktiga insikter om olika steg av det molekylära maskineriet. Viktiga mutationer identifieras och signalvägar undersöks utifrån gen- eller proteinuttryck.

En framgångssaga är användningen av Herceptin för HER2-positiv bröstcancer. HER2 är ett receptorprotein som sitter på cellmembranet, och detta protein är involverat i den signalöverföring som är viktig för cellernas överlevnad och utveckling i tumörerna. Herceptin är en molekyl som blockerar signaleringsprocessen genom detta receptorprotein och därigenom hindrar utvecklingen av de celler som uttrycker detta protein i hög grad. Ett annat lovande exempel är användningen av små molekyler som kallas tyrosin-kinas-hämmare, särskilt hos patienter som har en viss typ av cancer i buken (gastrointestinala stromacellstumörer).

I båda fallen riktas behandlingen mot ett protein av intresse, typiskt en svagpunkt i en känslig signalväg. Baserat på dessa principer är forskningen som presenteras i denna avhandling en samling av studier som syftar till att utöka vår förståelse av proteomet i olika typer av cancer. Artikel I beskriver en ny beräkningsmetod för att utvärdera reglering av proteinuttryck. Denna metod, och den mjukvara som implementerar metoden, ger prekliniska forskare möjlighet att identifiera de mekanismer som skiljer tumörceller från normala celler i motsvarande vävnad.

De återstående fyra artiklarna beskriver studier som fokuserar på proteomikanalys av olika former av cancer. I artiklarna II och III är kliniska prover från mjukdels sarkom och gastroesofageal tumörer analyserade och nya insikter om dessa komplexa sjukdomar med hjälp av viktiga proteiner och signalvägar läggs fram. Artikel IV och V är däremot fokuserade på bröstcancerbiologi. Artikel IV visar att det finns liten till ingen korrelation mellan proteinuttryck av ”odödliga” cellinjer och molekylära subtyper av bröstcancer, som dessa cellinjer förmodas representera. Resultaten från denna studie pekar mot att försiktighet måste iakttas vid överföring av kunskap från studier av dessa modellsystem. Artikel V presenterar en studie kring de terapeutiska aspekterna av ER-positiv, luminal bröstcancer. Vi har i synnerhet undersökt den systemiska betydelsen av östrogensignaleringsvägen och den terapeutiska potentialen av inhibering av denna väg demonstreras.

Det skulle vara alltför optimistiskt att ge en uppskattning på hur många år det kommer att ta innan vi har effektiva behandlingsmetoder mot alla former av cancer. Däremot är det klart att vi kommer att se viktiga förbättringar inom cancerterapi under de kommande 15-20 åren som i sin tur kommer att leda till förbättringar för patientens livskvalitet och ökad överlevnad.

Acknowledgments

As this long journey draws towards its conclusion, I cannot help but feel somewhat like what I imagine Christopher Columbus must have felt like; exhausted but proud, though not for having found a new continent but rather for having survived the oceans despite being lost over and over again, ending up somewhere so far away and foreign compared to what he could possibly be expecting... It has been a long journey, indeed, for more than 12 years I have been out there trying to make my way through the stormy and sunny days alike. At a moment like this you can't help but remember all the amazing and helpful people with whom you cross paths along the way.

I was born at a place where, unlike Barrack Obama's famous catch phrase, you are often told: "*No, you can't!*" I often heard that phrase even after I was out of the system that prevailed my homeland. Based on where I was born, my nationality (and probably my appearance) I was often told I could not get the things I wanted, which could very well explain my particularly sunny disposition in life. That being said, I was lucky enough to have met quite a few people who gave a helping hand, disregarding the general trend. Staring with the summer lab internship at Anders Björklund's lab, followed by another summer internship at Søren Brunak's lab in DTU with Nikolaj Blom, and finally a year-long project work again at the Wallenberg Neuroscience Center in Lund… My life seemed to be on a completely different track than what I could have imagined as a 15 year old, sitting in a plane bound for Copenhagen, traveling internationally for the first time in my life. A couple of years later I was enrolled at LTH, studying a program called Engineering Mathematics (!?), thanks to a fiercely intelligent but often silent man named Gunnar Sparr. It goes without saying that I had been originally told that my high-school education was poor, and that I could not study anything remotely technical or medical in Sweden, unless I re-did high school at Komvux. Ironically, they had not even bothered to check what my high school education looked like or how my grades were… Had it not been for the open-minded people like Anders, Nikolaj and Gunnar, I don't know where I would have been now, but I reckon it would certainly not be here, writing these words.

Fast forward about a decade... Here I am finishing up the last stitches of my doctoral thesis. This thesis would not have been a reality had it not been for my two supervisors, Peter and Fredrik.

Peter; thank you for being more of a teacher than a boss, thank you for showing that you can have a good sense of humour and be a complete guru in your field at the same time. Towards the second half of my graduate studies, I had realized that things were not going the way I had hoped, and that my thesis would not turn out as I had envisioned before I started. Then I came to remember that you had warned me (more than once actually) that this would be the likely case. Well, thanks once again, for not giving up on me.

Fredrik; thank you for always having your door open for me, no matter how many times I had bugged you already that day. I am amazed how you manage to have infinitely many "couple of minutes"! Don't know how I would do half of the things I have done if it wasn't for your quick fixes or code snippets.

I've had a couple of mentors over the years: Johan M, thanks for all our insightful conversations through out the years, I really appreciate you taking your time. I hope we can continue with our talks every now and then. Lars Erik, thanks for taking your time, sharing your extensive experience in life science research, for giving me the opportunity to get to know people way beyond my reach otherwise and incredibly expanding my network.

Heart-felt thanks for all colleagues, past and present, at "IT"; it's been fun to work you all. A couple of special mentions: Karin, thanks a lot for all your help and positive energy with the STS and Eso papers, honestly none of it would have been possible if you haven't helped me out with making some sense out of the experiments. Sofia, likewise thanks a lot for the fun and valuable talks, all the fantastic tips you gave for Manhattan, and for giving feedback on the first drafts of the thesis. Marianne, thanks for your energy that you dragged into the office every day and always having a minute or two for my (often silly) questions. Teleman, thanks for all off-the-books collaboration, in the form of discussions over our "rebel fika" 5-10 minutes before or after everybody else in the mornings. You are an intelligent man with a curious but disciplined mind, I have no doubt that you will be successful no matter you give yourself to after your PhD. Ola, although we never really worked together, it was really fun to "work" with you. I'm expecting a reunion in Japan! Paolo, sorry but Italy is so over-rated it's not like they have good wine, or cheese or ham... Thanks for everything, mate, the initial guidance and feedback in the later years, as well as the mutant frogs and smuggled beer glasses. Speaking of Italians, Vale, thanks for all the great banter, fantastic food and drinks. Anna S & Co for all sports-related and social events, fun to be around likeminded people.

No matter who you are, without a bunch of good friends you can't go far in life. I am really glad that I have had the fortune of meeting so many people. Kotte, Oscar, Kj, David, Emil… we have had our fair share of *bromance* over the years, thanks for being there both in good days and bad. Friends from university years (F-sek, E-sek and TLTH), friends from BMC D11, friends from MV Bld. 404; you are so many that I am afraid I would run out space and yet forget many good friends if I were to start listing names. I hope we will keep contact over the years, no matter where in the world life takes us. Anni, thanks for all the good memories, you have been a significant part of my life as a doctoral student. I hope things work out well for you in the end.

Last but least, my family... Bir insanin ilk okulu evi ve ailesidir. Yirmiye yakin yildir oyle veya boyle "okuyan" biri olarak ne mutlu bana boyle bir ailenin "kazan dibi" olarak dunyaya geldim. Canim annem ve babam, bana sadece matematik, tarih, cografya vs derslerimde degil ayni zamanda insani ve toplumsal degerleri, herhangi bir fikri, kavrami veya inanci empoze etmeden, kafa yikamadan ogrettiginiz icin size ne kadar tesekkur etsem de yetmez. Biz eger ulke sartlarinda standart bir aile olsaydiniz ne ben burada olurdum, ne de bu tez yazilmis olurdu.

Dino hocam, bir cay koysaydin ya... ☺

"It is not the critic who counts; not the man who points out how the strong man stumbles, or where the doer of deeds could have done them better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood; who strives valiantly; who errs, who comes short again and again, because there is no effort without error and shortcoming; but who does actually strive to do the deeds; who knows great enthusiasms, the great devotions; who spends himself in a worthy cause; who at the best knows in the end the triumph of high achievement, and who at the worst, if he fails, at least fails while daring greatly, so that his place shall never be with those cold and timid souls who neither know victory nor defeat. "

Theodore Roosevelt, *Citizens in a Republic*, Paris 1910

References

Aebersold, Ruedi. 2009. “A Stress Test for Mass Spectrometry-Based Proteomics..” *Nature Methods* 6 (6): 411–12. doi:10.1038/nmeth.f.255.

Aebersold, Ruedi, and Matthias Mann. 2003. “Mass Spectrometry-Based Proteomics.” *Nature* 422 (6928): 198–207. doi:10.1038/nature01511.

Anderson, N Leigh, and Norman G Anderson. 2002. “The Human Plasma Proteome: History, Character, and Diagnostic Prospects..” *Molecular & Cellular Proteomics : MCP* 1 (11): 845–67.

Anderson, N Leigh, Norman G Anderson, Lee R Haines, Darryl B Hardie, Robert W Olafson, and Terry W Pearson. 2004. “Mass Spectrometric Quantitation of Peptides and Proteins Using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA)..” *Journal of Proteome Research* 3 (2): 235–44.

Babur, Özgün, Ugur Dogrusoz, Emek Demir, and Chris Sander. 2010. “ChiBE: Interactive Visualization and Manipulation of BioPAX Pathway Models..” *Bioinformatics* 26 (3): 429–31. doi:10.1093/bioinformatics/btp665.

Backes, Christina, Andreas Keller, Jan Kuentzer, Benny Kneissl, Nicole Comtesse, Yasser A Elnakady, Rolf Müller, Eckart Meese, and Hans-Peter Lenhof. 2007. “GeneTrail--Advanced Gene Set Enrichment Analysis..” *Nucleic Acids Research* 35 (Web Server issue): W186–92. doi:10.1093/nar/gkm323.

Baird, Kristin, Sean Davis, Cristina R Antonescu, Ursula L Harper, Robert L Walker, Yidong Chen, Arthur A Glatfelter, Paul H Duray, and Paul S Meltzer. 2005. “Gene Expression Profiling of Human Sarcomas: Insights Into Sarcoma Biology..” *Cancer Research* 65 (20): 9226–35. doi:10.1158/0008-5472.CAN-05-1699.

Bantscheff, Marcus, Simone Lemeer, Mikhail M. Savitski, and Bernhard Kuster. 2012. “Quantitative Mass Spectrometry in Proteomics: Critical Review Update From 2007 to the Present..” *Analytical and Bioanalytical Chemistry* 404 (4). Springer-Verlag: 939–65. doi:10.1007/s00216-012-6203-4.

Beatson, GeorgeThomas. 1896. “On the Treatment of Inoperable Cases of Carcinoma of the Mamma: Suggestions for a New Method of Treatment, with Illustrative Cases..” *The Lancet* 148 (3803): 162–65. doi:10.1016/S0140-6736(01)72384-7.

Beck, Martin, Alexander Schmidt, Johan Malmstroem, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, and Ruedi Aebersold. 2011. “The Quantitative Proteome of a Human Cell Line..” *Molecular Systems Biology* 7: 549. doi:10.1038/msb.2011.82.

Bergamaschi, Anna, Geir Olav Hjortland, Tiziana Triulzi, Therese Sorlie, Hilde Johnsen, Anne Hansen Ree, Hege Giercksky Russnes, et al. 2009. "Molecular Profiling and Characterization of Luminal-Like and Basal-Like in Vivo Breast Cancer Xenograft Models.." *Molecular Oncology* 3 (5-6): 469–82. doi:10.1016/j.molonc.2009.07.003.

Bergman, O, G Hont, and E Johansson. 2013. *Cancer I Siffror 2013*. The National Board of Health and Welfare.

Bilbao, Aivett, Emmanuel Varesio, Jeremy Luban, Caterina Strambio-De-Castillia, Gérard Hopfgartner, Markus Müller, and Frederique Lisacek. 2015. "Processing Strategies and Software Solutions for Data-Independent Acquisition in Mass Spectrometry.." *Proteomics* 15 (5-6): 964–80. doi:10.1002/pmic.201400323.

Calligaris, David, Claude Villard, and Daniel Lafitte. 2011. "Advances in Top-Down Proteomics for Disease Biomarker Discovery.." *Journal of Proteomics* 74 (7): 920–34. doi:10.1016/j.jprot.2011.03.030.

Cancer Genome Atlas Network. 2012. "Comprehensive Molecular Portraits of Human Breast Tumours.." *Nature* 490 (7418): 61–70. doi:10.1038/nature11412.

Capriotti, Anna Laura, Chiara Cavaliere, Patrizia Foglia, Roberto Samperi, and Aldo Laganà. 2011. "Intact Protein Separation by Chromatographic and/or Electrophoretic Techniques for Top-Down Proteomics.." *Journal of Chromatography. A* 1218 (49): 8760–76. doi:10.1016/j.chroma.2011.05.094.

Carrillo, Brian, Corey Yanofsky, Sylvie Laboissiere, Robert Nadon, and Robert E Kearney. 2010. "Methods for Combining Peptide Intensities to Estimate Relative Protein Abundance.." *Bioinformatics* 26 (1). Oxford University Press: 98–103. doi:10.1093/bioinformatics/btp610.

Catherman, Adam D, Owen S Skinner, and Neil L Kelleher. 2014. "Top Down Proteomics: Facts and Perspectives.." *Biochemical and Biophysical Research Communications* 445 (4): 683–93. doi:10.1016/j.bbrc.2014.02.041.

Cerami, Ethan G, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. 2011. "Pathway Commons, a Web Resource for Biological Pathway Data.." *Nucleic Acids Research* 39 (Database issue): D685–90. doi:10.1093/nar/gkq1039.

Chapman, John D, David R Goodlett, and Christophe D Masselon. 2014. "Multiplexed and Data-Independent Tandem Mass Spectrometry for Global Proteome Profiling.." *Mass Spectrometry Reviews* 33 (6): 452–70. doi:10.1002/mas.21400.

Chawade, Aakash, Erik Alexandersson, and Fredrik Levander. 2014. "Normalyzer: a Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets.." *J. Proteome Res.* 13 (6). American Chemical Society: 3114–20. doi:10.1021/pr401264n.

Chawade, Aakash, Marianne Sandin, Johan Teleman, Johan Malmström, and Fredrik Levander. 2015. "Data Processing Has Major Impact on the Outcome of Quantitative Label-Free LC-MS Analysis.." *J. Proteome Res.* 14 (2): 676–87. doi:10.1021/pr500665j.

Claassen, Manfred. 2012. “Inference and Validation of Protein Identifications..” *Molecular & Cellular Proteomics* 11 (11). American Society for Biochemistry and Molecular Biology: 1097–1104. doi:10.1074/mcp.R111.014795.

CRICK, F H. 1958. “On Protein Synthesis..” *Symposia of the Society for Experimental Biology* 12: 138–63.

Demir, Emek, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, et al. 2010. “The BioPAX Community Standard for Pathway Data Sharing.” *Nature Biotechnology* 28 (9): 935–42. doi:10.1038/nbt.1666.

Deutsch, Eric W. 2011. “Tandem Mass Spectrometry Spectral Libraries and Library Searching..” *Methods Mol. Biol.* 696 (Chapter 13). Totowa, NJ: Humana Press: 225–32. doi:10.1007/978-1-60761-987-1_13.

Domon, Bruno, and Ruedi Aebersold. 2006. “Mass Spectrometry and Protein Analysis..” *Science* 312 (5771): 212–17. doi:10.1126/science.1124619.

Edman, P. 1950. “Method for Determination of the Amino Acid Sequence in Peptides.” *Acta Chem Scand.*

Eliuk, Shannon, and Alexander Makarov. 2015. “Evolution of Orbitrap Mass Spectrometry Instrumentation..” *Annual Review of Analytical Chemistry (Palo Alto, Calif.)* 8 (1): 61–80. doi:10.1146/annurev-anchem-071114-040325.

Emmert-Streib, Frank, and Galina V Glazko. 2011. “Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases.” Edited by Fran Lewitter. *PLOS Comput. Biol.* 7 (5): e1002053. doi:10.1371/journal.pcbi.1002053.t001.

Eng, J K, A L McCormack, and J R Yates. 1994. “An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database..” *Journal of the American Society for Mass Spectrometry* 5 (11). Springer-Verlag: 976–89. doi:10.1016/1044-0305(94)80016-2.

Fagerberg, Linn, Björn M Hallström, Per Oksvold, Caroline Kampf, Dijana Djureinovic, Jacob Odeberg, Masato Habuka, et al. 2014. “Analysis of the Human Tissue-Specific Expression by Genome-Wide Integration of Transcriptomics and Antibody-Based Proteomics..” *Molecular & Cellular Proteomics* 13 (2). American Society for Biochemistry and Molecular Biology: 397–406. doi:10.1074/mcp.M113.035600.

Fenn, J, M Mann, C Meng, S Wong, and C Whitehouse. 1989. “Electrospray Ionization for Mass Spectrometry of Large Biomolecules.” *Science* 246 (4926): 64–71. doi:10.1126/science.2675315.

Ferguson, Peter C, Benjamin M Deheshi, Peter Chung, Charles N Catton, Brian O'Sullivan, Abha Gupta, Anthony M Griffin, and Jay S Wunder. 2011. “Soft Tissue Sarcoma Presenting with Metastatic Disease: Outcome with Primary Surgical Resection..” *Cancer* 117 (2). Wiley Subscription Services, Inc., A Wiley Company: 372–79. doi:10.1002/cncr.25418.

Ferlay, Jacques, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. 2015. “Cancer

Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012..” *International Journal of Cancer* 136 (5): E359–86. doi:10.1002/ijc.29210.

Forshed, Jenny, Henrik J Johansson, Maria Pernemalm, Rui M M Branca, Annsofi Sandberg, and Janne Lehtiö. 2011. “Enhanced Information Output From Shotgun Proteomics Data by Protein Quantification and Peptide Quality Control (PQPQ)..” *Molecular & Cellular Proteomics* 10 (10): M111.010264. doi:10.1074/mcp.M111.010264.

Frantzi, Maria, Akshay Bhat, and Agnieszka Latosinska. 2014. “Clinical Proteomic Biomarkers: Relevant Issues on Study Design & Technical Considerations in Biomarker Development..” *Clinical and Translational Medicine* 3 (1). Springer: 7. doi:10.1186/2001-1326-3-7.

Fridley, Brooke L, Gregory D Jenkins, and Joanna M Biernacka. 2010. “Self-Contained Gene-Set Analysis of Expression Data: an Evaluation of Existing and Novel Methods.” Edited by Arkady B Khodursky. *PLoS ONE* 5 (9): e12693. doi:10.1371/journal.pone.0012693.t002.

Futcher, B, G I Latter, P Monardo, C S McLaughlin, and J I Garrels. 1999. “A Sampling of the Yeast Proteome..” *Molecular and Cellular Biology* 19 (11): 7357–68.

Geiger, Tamar, Juergen Cox, Pawel Ostasiewicz, Jacek R Wisniewski, and Matthias Mann. 2010. “Super-SILAC Mix for Quantitative Proteomics of Human Tumor Tissue.” *Nature Methods* 7 (5): 383–85. doi:10.1038/nmeth.1446.

Glazko, Galina V, and Frank Emmert-Streib. 2009. “Unite and Conquer: Univariate and Multivariate Approaches for Finding Differentially Expressed Gene Sets..” *Bioinformatics* 25 (18). Oxford University Press: 2348–54. doi:10.1093/bioinformatics/btp406.

Goeman, J J, and P Buhlmann. 2007. “Analyzing Gene Expression Data in Terms of Gene Sets: Methodological Issues.” *Bioinformatics* 23 (8): 980–87. doi:10.1093/bioinformatics/btm051.

Gouw, Joost W, Jeroen Krijgsveld, and Albert J R Heck. 2010. “Quantitative Proteomics by Metabolic Labeling of Model Organisms..” *Molecular & Cellular Proteomics* 9 (1). American Society for Biochemistry and Molecular Biology: 11–24. doi:10.1074/mcp.R900001-MCP200.

Gstaiger, Matthias, and Ruedi Aebersold. 2009. “Applying Mass Spectrometry-Based Proteomics to Genetics, Genomics and Network Biology.” *Nature Reviews Genetics* 10 (9): 617–27. doi:10.1038/nrg2633.

Guijarro, Maria V, Sonika Dahiya, Laura S Danielson, Miguel F Segura, Frances M Vales-Lara, Silvia Menendez, Dorota Popiolek, et al. 2013. “Dual Pten/Tp53 Suppression Promotes Sarcoma Progression by Activating Notch Signaling..” *The American Journal of Pathology* 182 (6): 2015–27. doi:10.1016/j.ajpath.2013.02.035.

Gygi, S P, B Rist, S A Gerber, F Turecek, M H Gelb, and R Aebersold. 1999. “Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags..” *Nature Biotechnology* 17 (10): 994–99. doi:10.1038/13690.

Hanahan, D, and R A Weinberg. 2000. “The Hallmarks of Cancer..” *Cell* 100 (1): 57–70.

Hanahan, Douglas. 2014. “Rethinking the War on Cancer..” *Lancet* 383 (9916): 558–63. doi:10.1016/S0140-6736(13)62226-6.

Hanahan, Douglas, and Robert A Weinberg. 2011. “Hallmarks of Cancer: the Next Generation..” *Cell* 144 (5): 646–74. doi:10.1016/j.cell.2011.02.013.

Hanash, Samir M. 2011. “Why Have Protein Biomarkers Not Reached the Clinic?.” *Genome Medicine* 3 (10). BioMed Central Ltd: 66. doi:10.1186/gm282.

Hu, Zhiyuan, Cheng Fan, Daniel S Oh, J S Marron, Xiaping He, Bahjat F Qaqish, Chad Livasy, et al. 2006. “The Molecular Portraits of Breast Tumors Are Conserved Across Microarray Platforms..” *BMC Genomics* 7 (1). BioMed Central Ltd: 96. doi:10.1186/1471-2164-7-96.

Hucka, M, A Finney, H M Sauro, H. Bolouri, J C Doyle, H Kitano, A P Arkin, et al. 2003. “The Systems Biology Markup Language (SBML): a Medium for Representation and Exchange of Biochemical Network Models..” *Bioinformatics* 19 (4): 524–31.

Hüttenhain, Ruth, Johan Malmström, Paola Picotti, and Ruedi Aebersold. 2009. “Perspectives of Targeted Mass Spectrometry for Protein Biomarker Verification..” *Current Opinion in Chemical Biology* 13 (5-6): 518–25. doi:10.1016/j.cbpa.2009.09.014.

Jemal, Ahmedin, Freddie Bray, Melissa M Center, Jacques Ferlay, Elizabeth Ward, and David Forman. 2011. “Global Cancer Statistics..” *CA: a Cancer Journal for Clinicians* 61 (2): 69–90. doi:10.3322/caac.20107.

Jensen, Ole Nørregaard. 2004. “Modification-Specific Proteomics: Characterization of Post-Translational Modifications by Mass Spectrometry..” *Current Opinion in Chemical Biology* 8 (1): 33–41. doi:10.1016/j.cbpa.2003.12.009.

Johansson, Henrik, Malin Lindstedt, Ann-Sofie Albrekt, and Carl AK Borrebaeck. 2011. “A Genomic Biomarker Signature Can Predict Skin Sensitizers Using a Cell-Based in Vitro Alternative to Animal Tests..” *BMC Genomics* 12: 399. doi:10.1186/1471-2164-12-399.

Karas, M, D Bachmann, U Bahr, and F Hillenkamp. 1987. “Matrix-Assisted Ultraviolet Laser Desorption of Non-Volatile Compounds.” *International Journal of Mass Spectrometry and Ion Processes* 78 (September): 53–68. doi:10.1016/0168-1176(87)87041-6.

Kelleher, N L, R A Zubarev, K Bush, B Furie, B C Furie, F W McLafferty, and C T Walsh. 1999. “Localization of Labile Posttranslational Modifications by Electron Capture Dissociation: the Case of Gamma-Carboxyglutamic Acid..” *Analytical Chemistry* 71 (19): 4250–53.

Keller, Andreas, Christina Backes, and Hans-Peter Lenhof. 2007. “Computation of Significance Scores of Unweighted Gene Set Enrichment Analyses.” *BMC Bioinformatics* 8 (1): 290. doi:10.1186/1471-2105-8-290.

Kern, Scott E. 2012. “Why Your New Cancer Biomarker May Never Work: Recurrent Patterns and Remarkable Diversity in Biomarker Failures..” *Cancer Research* 72 (23). American Association for Cancer Research: 6097–6101. doi:10.1158/0008-5472.CAN-12-3232.

Khatri, Purvesh, Marina Sirota, and Atul J Butte. 2012. "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.." *PLOS Comput. Biol.* 8 (2). Public Library of Science: e1002375. doi:10.1371/journal.pcbi.1002375.

Khoury, George A, Richard C Baliban, and Christodoulos A Floudas. 2011. "Proteome-Wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot Database.." *Scientific Reports* 1 (September). doi:10.1038/srep00090.

Kim, Min-Sik, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, et al. 2014. "A Draft Map of the Human Proteome.." *Nature* 509 (7502): 575–81. doi:10.1038/nature13302.

Kolkman, Annemieke, Maurien M A Olsthoorn, Carola E M Heeremans, Albert J R Heck, and Monique Slijper. 2005. "Comparative Proteome Analysis of Saccharomyces Cerevisiae Grown in Chemostat Cultures Limited for Glucose or Ethanol.." *Molecular & Cellular Proteomics : MCP* 4 (1): 1–11. doi:10.1074/mcp.M400087-MCP200.

Kondo, Tadashi. 2014. "Inconvenient Truth: Cancer Biomarker Development by Using Proteomics.." *Biochim. Biophys. Acta* 1844 (5): 861–65. doi:10.1016/j.bbapap.2013.07.009.

Koshy, Mary, Natia Esiashvilli, Jerome C Landry, Charles R Thomas, and Richard H Matthews. 2004. "Multiple Management Modalities in Esophageal Cancer: Combined Modality Management Approaches.." *The Oncologist* 9 (2): 147–59.

Lander, Arthur D. 2010. "The Edges of Understanding.." *BMC Biology* 8 (1). BioMed Central Ltd: 40. doi:10.1186/1741-7007-8-40.

Laukens, Kris, Stefan Naulaerts, and Wim Vanden Berghe. 2015. "Bioinformatics Approaches for the Functional Interpretation of Protein Lists: From Ontology Term Enrichment to Network Analysis.." *Proteomics* 15 (5-6): 981–96. doi:10.1002/pmic.201400296.

Ma, Bin, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. 2003. "PEAKS: Powerful Software for Peptide De Novo Sequencing by Tandem Mass Spectrometry.." *Rapid Communications in Mass Spectrometry* 17 (20). John Wiley & Sons, Ltd.: 2337–42. doi:10.1002/rcm.1196.

Maciejewski, Henryk. 2014. "Gene Set Analysis Methods: Statistical Models and Methodological Differences.." *Briefings in Bioinformatics* 15 (4). Oxford University Press: 504–18. doi:10.1093/bib/bbt002.

Makarov, Alexander, Eduard Denisov, Alexander Kholomeev, Wilko Balschun, Oliver Lange, Kerstin Strupat, and Stevan Horning. 2006. "Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer.." *Analytical Chemistry* 78 (7). American Chemical Society: 2113–20. doi:10.1021/ac0518811.

Makarov, Alexander, Eduard Denisov, Oliver Lange, and Stevan Horning. 2006. "Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer.." *Journal of the American Society for Mass Spectrometry* 17 (7). Springer-Verlag: 977–82. doi:10.1016/j.jasms.2006.03.006.

Mariette, Christophe, Guillaume Piessen, and Jean-Pierre Triboulet. 2007. "[Is There Still a Role for Surgery in Esophageal Carcinoma in 2007?].." *Bulletin Du Cancer* 94 (1): 63–69.

Matushansky, Igor, Eva Hernando, Nicholas D Socci, Joslyn E Mills, Tulio A Matos, Mark A Edgar, Samuel Singer, Robert G Maki, and Carlos Cordon-Cardo. 2007. "Derivation of Sarcomas From Mesenchymal Stem Cells via Inactivation of the Wnt Pathway.." *The Journal of Clinical Investigation* 117 (11): 3248–57. doi:10.1172/JCI31377.

McDonald, John H, University of Delaware. 2009. *Handbook of Biological Statistics*.

Michalski, Annette, Juergen Cox, and Matthias Mann. 2011. "More Than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS.." *J. Proteome Res.* 10 (4): 1785–93. doi:10.1021/pr101060v.

Mitchell, Melanie. 1998. *An Introduction to Genetic Algorithms*. MIT Press.

Moestue, Siver, Beathe Sitter, Tone Frost Bathen, May-Britt Tessem, and Ingrid Susann Gribbestad. 2011. "HR MAS MR Spectroscopy in Metabolic Characterization of Cancer.." *Current Topics in Medicinal Chemistry* 11 (1): 2–26.

Mootha, Vamsi K, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, et al. 2003. "PGC-1alpha-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes.." *Nature Genetics* 34 (3): 267–73. doi:10.1038/ng1180.

Nam, Dougu, and Seon-Young Kim. 2008. "Gene-Set Approach for Expression Pattern Analysis.." *Briefings in Bioinformatics* 9 (3): 189–97. doi:10.1093/bib/bbn001.

Nesvizhskii, Alexey I, and Ruedi Aebersold. 2005. "Interpretation of Shotgun Proteomic Data: the Protein Inference Problem.." *Molecular & Cellular Proteomics : MCP* 4 (10). American Society for Biochemistry and Molecular Biology: 1419–40. doi:10.1074/mcp.R500012-MCP200.

Olsson, Niclas, Peter James, Carl A K Borrebaeck, and Christer Wingren. 2012. "Quantitative Proteomics Targeting Classes of Motif-Containing Peptides Using Immunoaffinity-Based Mass Spectrometry.." *Molecular & Cellular Proteomics* 11 (8). American Society for Biochemistry and Molecular Biology: 342–54. doi:10.1074/mcp.M111.016238.

Ong, Shao-En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. 2002. "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.." *Molecular & Cellular Proteomics : MCP* 1 (5): 376–86.

Pan, Qun, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. 2008. "Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.." *Nature Genetics* 40 (12): 1413–15. doi:10.1038/ng.259.

Patani, Neill, Lesley-Ann Martin, and Mitch Dowsett. 2013. "Biomarkers for the Clinical Management of Breast Cancer: International Perspective.." *International Journal of Cancer* 133 (1): 1–13. doi:10.1002/ijc.27997.

Pavlidis, Paul, Jie Qin, Victoria Arango, John J Mann, and Etienne Sibille. 2004. “Using the Gene Ontology for Microarray Data Mining: a Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex..” *Neurochemical Research* 29 (6): 1213–22.

Perou, C M, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, et al. 2000. “Molecular Portraits of Human Breast Tumours..” *Nature* 406 (6797): 747–52. doi:10.1038/35021093.

Picotti, Paola, and Ruedi Aebersold. 2012. “Selected Reaction Monitoring-Based Proteomics: Workflows, Potential, Pitfalls and Future Directions..” *Nature Methods* 9 (6): 555–66. doi:10.1038/nmeth.2015.

Picotti, Paola, Bernd Bodenmiller, Lukas N. Mueller, Bruno Domon, and Ruedi Aebersold. 2009. “Full Dynamic Range Proteome Analysis of S. Cerevisiae by Targeted Proteomics..” *Cell* 138 (4): 795–806. doi:10.1016/j.cell.2009.05.051.

Picotti, Paola, Ruedi Aebersold, and Bruno Domon. 2007. “The Implications of Proteolytic Background for Shotgun Proteomics..” *Molecular & Cellular Proteomics : MCP* 6 (9): 1589–98. doi:10.1074/mcp.M700029-MCP200.

Purvine, Samuel, Jason-Thomas Eppel, Eugene C Yi, and David R Goodlett. 2003. “Shotgun Collision-Induced Dissociation of Peptides Using a Time of Flight Mass Analyzer..” *Proteomics* 3 (6). WILEY-VCH Verlag: 847–50. doi:10.1002/pmic.200300362.

Ransohoff, David F. 2005. “Lessons From Controversy: Ovarian Cancer Screening and Serum Proteomics..” *Journal of the National Cancer Institute* 97 (4). Oxford University Press: 315–19. doi:10.1093/jnci/dji054.

Raven, Peter H. 2005. *Biology*. McGraw-Hill Higher Education.

Rifai, Nader, Michael A Gillette, and Steven A Carr. 2006. “Protein Biomarker Discovery and Validation: the Long and Uncertain Path to Clinical Utility..” *Nature Biotechnology* 24 (8): 971–83. doi:10.1038/nbt1235.

Ross, Philip L, Yulin N Huang, Jason N Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, et al. 2004. “Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents..” *Molecular & Cellular Proteomics : MCP* 3 (12): 1154–69. doi:10.1074/mcp.M400129-MCP200.

Sandin, Marianne, Johan Teleman, Johan Malmström, and Fredrik Levander. 2014. “Data Processing Methods and Quality Control Strategies for Label-Free LC-MS Protein Quantification..” *Biochim. Biophys. Acta* 1844 (1 Pt A): 29–41. doi:10.1016/j.bbapap.2013.03.026.

Sawyers, Charles L. 2008. “The Cancer Biomarker Problem..” *Nature* 452 (7187): 548–52. doi:10.1038/nature06913.

Säll, Anna, Fredrika Carlsson, Niclas Olsson, Christer Wingren, Mats Ohlin, Helena Persson, and Sofia Waldemarson. 2014. “AFFIRM--a Multiplexed Immunoaffinity Platform That Combines Recombinant Antibody Fragments and LC-SRM Analysis..” *J. Proteome Res.* 13 (12): 5837–47. doi:10.1021/pr500757s.

Schenker, T, and B Trüeb. 1998. “Down-Regulated Proteins of Mesenchymal Tumor Cells..” *Experimental Cell Research* 239 (1): 161–68. doi:10.1006/excr.1997.3896.

Scherl, Alexander. 2015. “Clinical Protein Mass Spectrometry..” *Methods (San Diego, Calif.)* 81 (June): 3–14. doi:10.1016/j.ymeth.2015.02.015.

Schreier, T, R R Friis, K H Winterhalter, and B Trüeb. 1988. “Regulation of Type VI Collagen Synthesis in Transformed Mesenchymal Cells..” *The Biochemical Journal* 253 (2): 381–86.

Schulze, Waltraud X., and Björn Usadel. 2010. “Quantitation in Mass-Spectrometry-Based Proteomics..” *Annual Review of Plant Biology* 61: 491–516. doi:10.1146/annurev-arplant-042809-112132.

Schwanhäusser, Björn, Manfred Gossen, Gunnar Dittmar, and Matthias Selbach. 2009. “Global Analysis of Cellular Protein Translation by Pulsed SILAC..” *Proteomics* 9 (1). WILEY-VCH Verlag: 205–9. doi:10.1002/pmic.200800275.

Serang, Oliver, and William Noble. 2012. “A Review of Statistical Methods for Protein Identification Using Tandem Mass Spectrometry..” *Statistics and Its Interface* 5 (1). NIH Public Access: 3–20. doi:10.1021/pr400678m.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks..” *Genome Research* 13 (11): 2498–2504. doi:10.1101/gr.1239303.

Shteynberg, David, Alexey I Nesvizhskii, Robert L Moritz, and Eric W Deutsch. 2013. “Combining Results of Multiple Search Engines in Proteomics..” *Molecular & Cellular Proteomics* 12 (9). American Society for Biochemistry and Molecular Biology: 2383–93. doi:10.1074/mcp.R113.027797.

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal. 2015. “Cancer Statistics, 2015.” *CA: a Cancer Journal for Clinicians* 65 (1): 5–29. doi:10.3322/caac.21254.

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. CRC Press.

Skrbo, Nirma, Geir Olav Hjortland, Alexandr Kristian, Ruth Holm, Silje Nord, Lina Prasmickaite, Olav Engebraaten, Gunhild M Maelandsmo, Therese Sorlie, and Kristin Andersen. 2014. “Differential in Vivo Tumorigenicity of Distinct Subpopulations From a Luminal-Like Breast Cancer Xenograft..” *PLoS ONE* 9 (11). Public Library of Science: e113278. doi:10.1371/journal.pone.0113278.

Sorlie, Therese, Robert Tibshirani, Joel Parker, Trevor Hastie, J S Marron, Andrew Nobel, Shibing Deng, et al. 2003. “Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets..” *Proceedings of the National Academy of Sciences of the United States of America* 100 (14): 8418–23. doi:10.1073/pnas.0932692100.

Stark, G R. 1968. “Sequential Degradation of Peptides From Their Carboxyl Termini with Ammonium Thiocyanate and Acetic Anhydride..” *Biochemistry* 7 (5): 1796–1807.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment

Analysis: a Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles..” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. doi:10.1073/pnas.0506580102.

Syka, John E P, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. 2004. “Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry..” *Proceedings of the National Academy of Sciences of the United States of America* 101 (26). National Acad Sciences: 9528–33. doi:10.1073/pnas.0402700101.

Szallasi, Zoltan, Jörg Stelling, and Vipul Periwal. 2010. *System Modeling in Cellular Biology*. MIT Press (MA).

Sørlie, T, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, et al. 2001. “Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications..” *Proceedings of the National Academy of Sciences of the United States of America* 98 (19): 10869–74. doi:10.1073/pnas.191367098.

Tabb, David L, Lorenzo Vega-Montoto, Paul A Rudnick, Asokan Mulayath Variyath, Amy-Joan L Ham, David M Bunk, Lisa E Kilpatrick, et al. 2010. “Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography-Tandem Mass Spectrometry..” *J. Proteome Res.* 9 (2): 761–76. doi:10.1021/pr9006365.

Tanaka, Koichi, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, and T Matsuo. 1988. “Protein and Polymer Analyses Up Tom/Z 100 000 by Laser Ionization Time-of-Flight Mass Spectrometry.” *Rapid Communications in Mass Spectrometry* 2 (8): 151–53. doi:10.1002/rcm.1290020802.

Tew, William P, David P Kelsen, and David H Ilson. 2005. “Targeted Therapies for Esophageal Cancer..” *The Oncologist* 10 (8): 590–601. doi:10.1634/theoncologist.10-8-590.

Thiele, Herbert, Stefan Heldmann, Dennis Trede, Jan Strehlow, Stefan Wirtz, Wolfgang Dreher, Judith Berger, et al. 2014. “2D and 3D MALDI-Imaging: Conceptual Strategies for Visualization and Data Mining..” *Biochim. Biophys. Acta* 1844 (1 Pt A): 117–37. doi:10.1016/j.bbapap.2013.01.040.

Thompson, Andrew, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, R Johnstone, A Karim A Mohammed, and Christian Hamon. 2003. “Tandem Mass Tags: a Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS..” *Analytical Chemistry* 75 (8): 1895–1904.

UniProt Consortium. 2015. “UniProt: a Hub for Protein Information..” *Nucleic Acids Research* 43 (Database issue). Oxford University Press: D204–12. doi:10.1093/nar/gku989.

van de Rijn, Matt, and Jonathan A Fletcher. 2006. “Genetics of Soft Tissue Tumors..” *Annual Review of Pathology* 1: 435–66. doi:10.1146/annurev.pathol.1.110304.100052.

Venable, John D, Meng-Qiu Dong, James Wohlschlegel, Andrew Dillin, and John R Yates. 2004. “Automated Approach for Quantitative Analysis of Complex Peptide Mixtures From Tandem Mass Spectra..” *Nature Methods* 1 (1): 39–45. doi:10.1038/nmeth705.

Walsh, Christopher. 2006. *Posttranslational Modification of Proteins*. Roberts and Company Publishers.

Wasinger, V C, S J Cordwell, A Cerpa-Poljak, J X Yan, A A Gooley, M R Wilkins, M W Duncan, R Harris, K L Williams, and I Humphery-Smith. 1995. "Progress with Gene-Product Mapping of the Mollicutes: Mycoplasma Genitalium.." *Electrophoresis* 16 (7): 1090–94.

Webb-Robertson, Bobbie-Jo M, Melissa M Matzke, Susmita Datta, Samuel H Payne, Jiyun Kang, Lisa M Bramer, Carrie D Nicora, et al. 2014. "Bayesian Proteoform Modeling Improves Protein Quantification of Global Proteomic Measurements.." *Molecular & Cellular Proteomics* 13 (12). American Society for Biochemistry and Molecular Biology: 3639–46. doi:10.1074/mcp.M113.030932.

Weinberg, Robert. 2013. *The Biology of Cancer, Second Edition*. Garland Science.

Werner, Thilo, Gavain Sweetman, Maria Fälth Savitski, Toby Mathieson, Marcus Bantscheff, and Mikhail M. Savitski. 2014. "Ion Coalescence of Neutron Encoded TMT 10-Plex Reporter Ions.." *Analytical Chemistry* 86 (7): 3594–3601. doi:10.1021/ac500140s.

Wilhelm, Mathias, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, et al. 2014. "Mass-Spectrometry-Based Draft of the Human Proteome.." *Nature* 509 (7502): 582–87. doi:10.1038/nature13319.

Wilkins, M R, J C Sanchez, A A Gooley, R D Appel, I Humphery-Smith, D F Hochstrasser, and K L Williams. 1996. "Progress with Proteome Projects: Why All Proteins Expressed by a Genome Should Be Identified and How to Do It.." *Biotechnology & Genetic Engineering Reviews* 13: 19–50.

Wolf-Yadlin, Alejandro, Sampsa Hautaniemi, Douglas A Lauffenburger, and Forest M White. 2007. "Multiple Reaction Monitoring for Robust Quantitative Proteomic Analysis of Cellular Signaling Networks.." *Proceedings of the National Academy of Sciences of the United States of America* 104 (14): 5860–65. doi:10.1073/pnas.0608638104.

Yates, John R, Cristian I Ruse, and Aleksey Nakorchevsky. 2009. "Proteomics by Mass Spectrometry: Approaches, Advances, and Applications.." *Annual Review of Biomedical Engineering* 11: 49–79. doi:10.1146/annurev-bioeng-061008-124934.

Zubarev, R A, D M Horn, E K Fridriksson, N L Kelleher, N A Kruger, M A Lewis, B K Carpenter, and F W McLafferty. 2000. "Electron Capture Dissociation for Structural Characterization of Multiply Charged Protein Cations.." *Analytical Chemistry* 72 (3): 563–73.