LUND UNIVERSITY

# Building a map of the breast cancer proteome - Strategies to increase coverage

Cifani, Paolo

2013

Link to publication

*Citation for published version (APA):*
Cifani, P. (2013). *Building a map of the breast cancer proteome - Strategies to increase coverage*. [Doctoral Thesis (compilation), Department of Immunotechnology].

*Total number of authors:*
1

**General rights**
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

**Take down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

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

Download date: 07. Jul. 2025

# Building a map of the breast cancer proteome

## Strategies to increase coverage


LUNDS UNIVERSITET

Paolo Cifani

ACADEMIC THESIS,
which by due permission of the
Faculty of Engineering, Lund University, Sweden,
will be publicly defended at Hörsalen, Medical Village, Scheelevägen 2, Lund,
on Friday the 31st of May 2013 at 9.15 .

Faculty opponent is Prof. Luca Bini,
Dipartimento di Biologia Molecolare, Università di Siena, Italy.

Organization
LUND UNIVERSITY
Department of Immunotechnology
Medicon Village (H406)
SE-223 81 Lund
Sweden

Document name
DOCTORAL DISSERTATION
Date of issue
2013-05-31

Author(s)
Paolo Cifani

Sponsoring organization

Title and subtitle
Building a map of the breast cancer proteome
Strategies to increase coverage

Abstract
Amongst the various –omics sciences, proteomics has the highest potential for functional characterization and consequently can contribute significantly to the field of cancer research. In particular, the focus of this thesis is on breast cancer. Alas, since state-of-the-art technologies cannot meet the complexity of upper eukaryotic proteomes, a complete resolution of clinical samples is still unachievable. Comprehensive mapping of proteins involved in cancer and of their PTMs is proposed in this thesis as a general strategy to increase the output of mass-spectrometry based proteomics. Different approaches to improve the coverage of this map are proposed: optimization of sample fractionation, focusing on difficult sub-proteomes, targeting of specific biological processes and optimization of data analysis. A combination of these approaches will provide a growing collection of empirical MS-spectra, which will enhance the detection by shotgun proteomics and facilitate the transition towards the development of targeted assays.

Key words
Quantitative proteomics, Mass-Spectrometry, Cancer, DNA repair, Hypoxia, PTMs, Sample fractionation, Database

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-7473-529-1
978-91-7473-530-7 (pdf)

Recipient´s notes

Number of pages

Price

Security classification

Distribution by (name and address)

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

Signature Paolo Cifani Date 2013-04-22

*There is and will remain a Platonic element in science which could not be taken away without ruining it. Among the infinite diversity of singular phenomena science can only look for invariants*.

Jacques Monod

# TABLE OF CONTENTS

**Original papers** **6**

**Author's contribution to the papers** **8**

**Other papers** **9**

**Abbreviations** **10**

**Introduction** **11**

**Proteomics** **13**
*Mass spectrometry-based proteomics* *13*
Mass-spectrometers 14
Shotgun and targeted proteomics 15
*Divide et impera* *19*
*Quantitative shotgun proteomics* *21*
Gel-based quantitation 22
Label-free quantitation 22
Stable isotope labelling 23
SILAC 24
*Difficult sub-proteomes* *26*
Phosphoproteomics 26
Membrane proteomics 28
*Proteomics: a science in its teens?* *30*

**Cancer** **33**
*DNA repair in breast cancer* *35*
DNA damage and repair 38
Radiotherapy and DNA double-strand breaks repair 39
*Hypoxia in solid tumours: neuroblastoma* *44*
Tumour Hypoxia 44

**Cancer Proteomics** **47**
*Biomarkers* *47*
*Towards a proteomic definition of "The Hallmarks of Cancer"* *49*
*What was sampled?* *50*
*Predictably unpredictable* *52*
*Finding the way with a proteomic map* *53*

**Conclusions and future perspectives** **55**
*Proteomics as a grown-up science: which way to go?* *56*

**Populärvetenskaplig sammanfattning** **59**

**Summary** **61**

**Sintesi a scopo divulgativo** **63**

**Acknowledgements** **65**

**References** **67**

# ORIGINAL PAPERS

This thesis is based upon the following papers, which are referred to in the text by their Roman numerals (I-VI). The papers can be found as appendices at the end of the book. PAPER I, II and V are reprinted with permission from the publisher. Copyright 2011 (PAPER II) and 2012 (PAPER I and V), American Chemical Society.

PAPER I:
Antberg L.*, **Cifani P.***, Sandin M., Levander F., James P.
Critical comparison of multidimensional separation methods for increasing protein expression coverage.
*J Proteome Res*. 2012 May 4;11(5):2644-52.

PAPER II:
**Cifani P.***, Bendz M.*, Wårell K., Hansson K., Levander F., Sandin M., Krogh M., Ovenberger M., Fredlund E., Vaapil M., Pietras A., Påhlman S., James P.
Hunting for protein markers of hypoxia by combining plasma membrane enrichment with a new approach to membrane protein analysis.
*J Proteome Res. 2011 Apr 1;10(4):1645-56.*

PAPER III
**Cifani P.**, Kirik U., James P.
Indexing the breast proteome: a portrait of five commonly used breast cancer cell lines.
Manuscript submitted for publication

PAPER IV
**Cifani P**., James P.,
A proteomic analysis of radio-resistance in breast cancer cell lines.
Manuscript in preparation

PAPER V
Kirik U., **Cifani P.**, Albrekt A.S., Lindstedt M., Heyden A., Levander F..
Multimodel pathway enrichment methods for functional evaluation of expression regulation.
*J Proteome Res. 2012 May 4;11(5):2955-67.*

PAPER VI
Antberg L., **Cifani P.**, Levander F., James P.
Pathway-centric analysis of the DNA damage response to chemotherapeutic agents in two breast cell lines.
Manuscript submitted for publication

*= these authors equally contributed to the manuscript.

# AUTHOR'S CONTRIBUTION TO THE PAPERS

PAPER I:
Together with LA, PC designed and performed the experiments, participated in data analysis and in writing the manuscript.

PAPER II:
PC analyzed the data and drafted the manuscript

PAPER III:
PC designed and performed all experiments, executed most of the data analysis and wrote the manuscript.

PAPER IV:
PC designed and performed the experiments and the data analysis and wrote the manuscript.

PAPER V:
PC contributed to the conceptual design of the software.

PAPER VI:
PCoptimized the SRM assays for the quantitation DNA repair enzymes.

# OTHER PAPERS

The author also took part in the following work, not included in this thesis:

Stella R., **Cifani P.**, Peggion C., Hansson K., Lazzari C., Bendz M., Levander F., Sorgato M.C., Bertoli A., James P.
Relative quantification of membrane proteins in wild-type and prion protein (PrP)-knockout cerebellar granule neurons.
*J Proteome Res. 2012 Feb 3;11(2):523-36.*

# ABBREVIATIONS

| | |
|---|---|
| AIF: | all-ion fragmentation. |
| CAD: | collisionally activated dissociation. |
| CID: | collision-induced acquisition. |
| DDA: | data dependent acquisition. |
| DIA: | data independent acquisition. |
| DIGE: | difference gel electrophoresis. |
| ESI: | electro-spray ionization. |
| ETD: | electron transfer dissociation. |
| IEF: | isoelectric focusing. |
| IT: | ion trap. |
| HCD: | higher-energy collisional dissociation. |
| HER+: | HER over-expressing. |
| LC: | liquid chromatography. |
| MALDI: | matrix assisted laser desorption ionization. |
| MRM: | multiple reaction monitoring. |
| MS: | mass spectrometer / mass spectrometry. |
| NicNHS: | nicotinoyl-n-hydroxysuccinimide. |
| PACiFIC: | precursor acquisition independent from ion count. |
| PAGE: | polyacrylamide gel electrophoresis. |
| PEG: | polyethylene glycol. |
| PTM: | post-translational modification. |
| Q: | quadrupole. |
| ROS: | reactive oxygen species. |
| RP: | reverse phase. |
| SCX: | strong cation exchange. |
| SDS: | sodium dodecyl sulphate. |
| SILAC: | stable isotope labeling of amino acids in cell culture. |
| SRM: | selected reaction monitoring. |
| TMT: | tandem mass tag. |
| TSQ: | triple-stage quadrupole. |

# INTRODUCTION

In the past two decades, high-throughput methodologies for protein analysis have enjoyed a remarkable technological development, mostly based on the improvements in the performance of mass-spectrometers. As a result, identification and quantitation of thousands of proteins and of their chemical modifications is currently achievable with much higher efficiency compared to pre-existing methods. The development of such a powerful toolkit prompted its immediate application to one of the research fields where quick breakthroughs are most needed because of high incidence and mortality of the studied disease: breast cancer. This malignancy in fact, despite recent progresses in diagnosis and treatment, remains one of the leading causes of death for women.

The application of *–omics* approaches to cancer research did not start with proteomics. Genomic studies provided insights into the aetiology of cancer by revealing specific mutations associated with the disease and highlighting the connection between defective DNA repair and the onset of the malignancy. At the same time, these elements provided markers of increased cancer susceptibility, like BRCA mutations, and opened novel treatment strategies such as PARP inhibitor chemotherapics. Transcriptomic technologies on the other hand yielded the most comprehensive molecular classification scheme proposed so far. Since "breast cancer" is a rather broad definition encompassing many distinct pathologies, an objective system to differentiate between different tumour types enables a more targeted and effective choice of treatment.

The promise of a large-scale characterization of tumour proteins gathered high expectations around proteomics but concrete results with immediate beneficial effects on the clinical practice are still awaited. This does not mean that mass spectrometry-based proteomics is not suited to cancer research. The lack of clinically relevant discoveries indicates instead that technological improvements as well as conceptual adjustments are needed to promote the growth of proteomics into a "mature" science.

This thesis addresses how proteomics can contribute to breast cancer research and suggests strategies to improve the effectiveness of the related

technologies. One basic concept will drive the discussion: current proteomic techniques cannot cope with the complexity of entire cancer proteome but at lot can be gained by dividing it in discrete portions. The data included in these patches are then going to form a global map of cancer proteome. This map, rather than being the end-point of cancer proteomics, would constitute a framework to both drive further discovery oriented studies and facilitate translation of the proteomic findings into clinical practice.

The thesis first presents an overview of mass-spectrometry based proteomics and of some of its limitations. In consideration of the complexity of the topic, the discussion is mostly restricted to the aspects that are relevant for the papers presented or for future developments. The focus is mainly on shotgun proteomics with brief insights into the targeted approach.

The following chapter introduces elements of cancer biology, with special emphasis on processes related to tumour hypoxia and DNA repair. The discussion about cancer provides the context for the studies presented later and at the same time provides the rationale for increasing the resolution of specific portions of the "cancer proteome map". Breast cancer is the main topic of the studies but neuroblastoma is also presented as a model system for tumour hypoxia. Radiotherapy in breast cancer treatment is described in the context of DNA repair.

Having introduced both the technology and the biological problem, how proteomics could, or should, contribute to cancer research is finally discussed in "Cancer proteomics". Issues concerning biomarkers discovery and model systems are described and the concept of cancer proteomic map is examined.

The contribution of each paper to the topics above is briefly presented in the thesis along with the discussion and finally summarized in the last chapter together with future perspectives.

# PROTEOMICS

To functionally characterise a biological sample entails, oversimplifying a little, answering questions such as: "Which protein species exist in the sample?", "Are these proteins modified?" and "What is the amount of each protein?". The entire protein content in a sample, with this generic term encompassing anything from sub-cellular compartments to entire organisms, is named *proteome*. *Proteomics*, the system-wide study of the proteome, is therefore the science that addresses the questions above. Considering the holistic approach of proteomics, the tools employed in the discipline need to be high-throughput and capable of meeting the chemical heterogeneity, the variability over time and the wide range of concentrations found with proteins (Corthals et al. 2000; N. L. Anderson & N. G. Anderson 2002). Practically these tools can be grouped into two major families: antibody-based technologies, such as protein microarrays (Haab et al. 2001), and workflows relying on mass spectrometry (MS). Technological improvements over the past two decades have turned the latter discipline into a powerful approach to address the issues presented by proteomics.

## Mass spectrometry-based proteomics

Two main approaches for proteins MS analysis can be defined: *top-down* and *bottom-up*. In a top-down experiment, intact proteins are submitted to the MS that fragments the analytes and provides the *m/z* measurement for both the intact precursor and the corresponding fragments. This method is useful to identify proteins with high sequence coverage and to localize post-translation modifications and variations of the primary structure (Kelleher et al. 1999), but it presents limitations regarding the molecular mass of the analyte and is currently not suitable for complex samples. Conversely the bottom-up approach, on which most of this thesis is based and the rest of the discussion will be focused, consists on the digestion of sample proteins by mean of sequence-specific proteases such as trypsin (Olsen et al. 2004), followed by

tandem-MS analysis (MS/MS) of the resulting peptides (Link et al. 1999). Since the upfront separation of peptides is technically simpler compared to that of intact proteins, bottom-up proteomics allows the quick and straightforward identification of thousands of proteins (Thakur et al. 2011) but at same time it is affected by poorer sequence coverage, loss of labile post-translational modifications (PTMs) and the inability to assign non-unique peptides to their protein of origin.

## Mass-spectrometers

A mass spectrometer essentially consists of three parts; an ion source capable of transferring the analytes from the solid or liquid phase to the gas-phase, a mass analyser that separates the ions according to their *m/z* (mass over charge ratio) and a suitable ion detector (Aebersold & Mann 2003).

Bottom-up proteomics requires mass spectrometers capable of measuring the mass of the intact peptide (*precursor*), fragmenting it and determining the *m/z* values for the obtained *daughter ions*. ESI (*Electrospray Ionization*, Fenn et al. 1989) and MALDI (*Matrix Assisted Laser Desorption Ionization*, Karas & Hillenkamp 1988) are the most common sources since they are "soft" ionization methods (i.e. do not cause extensive fragmentation of the peptide). The first is based on desolvation of peptide-containing droplets nebulised by a capillary at high electrical potential compared to the MS. ESI is a continuous source, therefore suitable to be coupled with upfront chromatographic devices, and often produces multiply charged ions. MALDI is instead pulsed by nature and the ions generated are normally singly charged. The peptides to be ionized are embedded in a matrix that upon excitation by a laser beam aids sample desorption and ionization.

*Time-of-flight* (TOF) mass analyzers measure the flight time of molecules in a tube under vacuum and of known length. *Quadrupoles* (Q) select molecules with given *m/z* according to their stability in an oscillating radio-frequency electric field but the movement of the ions in the direction perpendicular to the field is not restricted. *Ion traps* (IT) can be thought of as quadrupoles in which the ions motility is controlled also in the third dimension, therefore allowing for accumulation of analytes over time. In the *orbitrap* mass analyzers (Hardman & Makarov 2003), ions are trapped between an inner and an outer coaxial electrodes and oscillate along the device axis with a frequency dependent on the specific *m/z*. Unlike ion traps, this

instrument allows non-destructive measurements of MS1 spectra and has the highest resolution between the mass analyzers presented here.

*Collision-Induced Dissociation* (CID, also known as *CAD, collisionally activated dissociation*) is probably the most widespread peptides fragmentation methods. In the fragmentation cell, the cations are kinetically excited by radiofrequency to collide with an inert gas such as Helium or Argon and as a result the backbone of the precursor peptides breaks producing two or more daughter ions (Hunt et al. 1986). *Higher-energy C-trap Dissociation* (HCD) exploits the same basic principle of CID but at higher frequency (Olsen et al. 2007). *Electron-Transfer Dissociation* (ETD) induces fragmentation independently of amide bond protonation by mean of ion/ion reaction of multiply charged peptide with an anionic reagent (Syka et al. 2004).

Different combinations of these components have created a relatively wide array of instruments with specific strengths and drawbacks, each suitable for different experimental requirements (Domon & Aebersold 2006). A linear IT with ESI source was used in PAPER I, where speed and sensitivity of the instrument were more important than high mass-accuracy. In PAPER II, a MALDI-TOF was used to support a top-down approach while a complex mixture of modified membrane proteins were resolved by an ESI-Q-TOF. Both PAPERS III, IV and VI took advantage of one of the preferred instrument for discovery proteomic: a hybrid ESI-LTQ-Orbitrap (Q. Hu et al. 2005; Scigelova & Makarov 2006). This instrument in fact features two mass-detectors: a fast and sensitive but low-resolution linear ion trap and the high-mass accuracy and high-resolution orbitrap itself. As a result, the ESI-LTQ-Orbitrap (from now on simply referred to as *orbitrap*) permits the slow acquisition of high resolution MS1 in parallel with fast collection of MS2. Finally it is worth mentioning the *triple (stage) quadrupole* (TSQ/QQQ), an instrument with very high sensitivity and dynamic range that is the workhorse of targeted quantitative proteomics described below. A TSQ consists of three quadrupoles ($Q_1$, $Q_2$ and $Q_3$): $Q_1$ and $Q_3$ are operated as mass-analyzers while $Q_2$ is the collision cell. TSQ was used in PAPER V to perform selected reaction monitoring (SRM) analysis.

## Shotgun and targeted proteomics

Depending on the experimental goals, the mass spectrometer can either be used to iteratively scan a given mass range and select the dominant ions for

fragmentation ("*shotgun*" proteomics, *Fig.1*) or to measure some pre-defined peptide-specific *m/z* values (*targeted* proteomics).

In the first case, the recorded tandem MS spectra can be decoded into a peptide sequence in three ways, the most common of which consists in matching the measured signals to predicted spectra obtained by *in silico* translation, digestion and fragmentation of the relevant genome (*database searching*, Steen & Mann 2004; Eng et al. 2011). This task is performed by search engines such as Mascot (Perkins et al. 1999) and, more recently, Andromeda (Jürgen Cox et al. 2011). Database search identification is a widely accepted tool but is hampered by the limited number of PTMs that one can include in every search and by the dependency of the retrieved results on the specific target database.

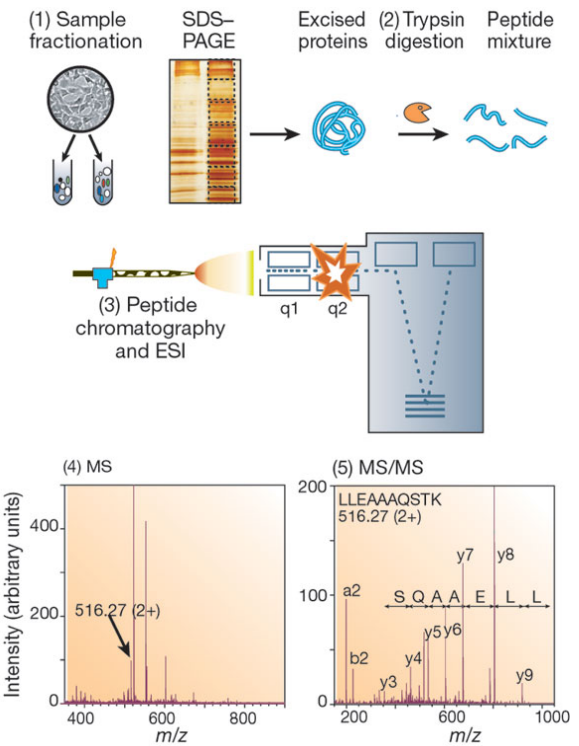


**Figure 1**
*The shotgun proteomic work-flow (Aebersold & Mann 2003).*

In a second option, to assign a measured fragmentation pattern to an amino acidic sequence the predicted spectra used in database searching is replaced by previously measured reference spectra stored in a *spectral library* (Craig et al. 2006; Lam 2011), This change of paradigm will be further discussed later in this thesis. The third approach to peptide identification, *de novo sequencing*, provides peptide sequences without using any previous database: the residues identity and their order in the precursor are inferred by the mere examination of the tandem spectrum, considering all the possible amino acid permutations (B. Ma & Johnson 2012).

Once the peptides sequences have been determined, these fragments are matched to the protein which had originated them in a similar way to the namesake genomic sequencing strategy, hence the denomination (Claassen 2012; Hoopmann & Moritz 2013). This process relies by necessity on genome-derived databases and it can be quite insidious. Firstly, peptides are sequenced with different confidence levels, therefore affecting the reliability of protein identification. Moreover, many peptides are *degenerate,* i.e. derive from more than one protein, complicating both identification and quantification (discussed later) of the protein (Huang et al. 2012). As an obvious corollary, pretending that every protein would have on average the same sequence coverage, shotgun proteomics identification and quantitation of bigger proteins is favoured compared to that of smaller species.

The shotgun strategy is generally adopted in the *discovery* phase of proteomics studies because in principle it does not require any prior knowledge about the sample composition, except for the taxonomy of the sampled organism when database searching is used.

In the majority of the experiments the MS is run in *data dependent acquisition* (DDA) mode, which consists of iteratively selecting the most abundant ions in a certain mass window for fragmentation. Thus, detection and quantitation of low-abundance proteins is often problematic. This aspect will be covered with more details later in this chapter. *Data Independent Acquisition* (DIA), i.e. the iterative analysis of discrete portions of the *m/z* range independent from the presence and intensity of chromatographic peaks (Venable et al. 2004), has been proposed to increase the detectability of peptides at the low-end of the dynamic range but so far this approach has found little application. An example of DIA workflow worth mentioning is the PAcIFIC (*Precursor Acquisition Independent From Ion Count)* strategy which proceeds selecting narrow and overlapping *m/z* ranges for CID fragmentation and using the centres of these windows as precursor mass for database searching (Panchaud et al. 2009). Geiger and co-worker described a full range

DIA (*All Ions Fragmentation*: AIF) using an orbitrap mass analyzer to obtain high resolution spectra of both precursor and daughter ions (Geiger, Juergen Cox & Mann 2010a).

Instead of sampling the peptide mixture in a stochastic manner, the MS can be set to detect only specific precursors and/or fragmentation products. This *targeted* approach can be performed on "scanning" instruments (*Selected Ions Monitoring:* SIM) but is mostly employed on TSQs, taking the name of *Selective Reaction Monitoring* (SRM) or *Multiple Reaction Monitoring* (MRM). The non-scanning nature of this MS strategy, where the first and the third quadrupoles of the instrument are used as filters to ensure selectivity on the precursor and on the daughter ions respectively, results in higher sensitivity and more accurate quantitation compared to standard shotgun strategies (Lange et al. 2008). This allows to reproducibly track the abundance of specific peptides across multiple samples but the need for pre-defined assays for every analyte of interest practically confines SRM and MRM to the post-discovery phase. This targeted workflow proved itself robust even across different laboratories (Addona et al. 2009), a very desirable feature in order to adapt the technique to a clinical setting. The critical step in the development of peptide-specific assays is the definition of the *m/z* values to be measured in the $Q_1$ and $Q_3$ (*transitions*), a task that is greatly facilitated by empirical spectral libraries like those used for spectral searching. Another factor limiting the throughput of SRM consists in the time employed by the TSQ to measure all the transitions in the list. Since each chromatographic peak needs to be profiled over time, the number of ions that can be tracked in each run is practically rather limited.

Targeted detection of proteins involved in DNA repair is described in PAPER VI. In this work SRM analysis proved itself capable of detecting and quantitating DNA-repair proteins, many of which are known to be expressed a relatively low abundance in un-fractionated lysate of human cells.

Combining DIA with a targeted-like data analysis, the SWATH approach (Gillet et al. 2012) shows the potential to overcome the low-throughput limitations of SRM. This new strategy is still under-development and its feasibility with complex proteomes such as those of higher eukaryotes is at present questionable, but in principle the SWATH analysis would blur the boundary between targeted and shotgun proteomics.

## *Divide et impera*

Whichever type of mass spectrometer one may consider, it will always have a finite scan rate (i.e. the number of measurements that the instrument can perform per time unit). This means that, as soon as the number of ions fed to the instrument per time unit exceeds its scanning capability, under-sampling will cause some ions to remain undetected. In the context of shotgun experiments, this situation favours the measurement of the most abundant or more readily ionisable peptides at the expense of scarcer ions (H. Liu et al. 2004). Therefore, keeping the MS specifications (scan rate, resolution, sensitivity) constant, the number of peptides successfully sequenced is largely dependent on the efficacy of the upfront sample separation. It follows that the use of pulsed ion sources (for example MALDI) is restricted to relatively simple (or resolved) samples, while the analysis of complex mixtures requires peptides to be separated and delivered to the MS in a time-delayed fashion, which in practical terms is achieved by using a continuous ESI source in line with a chromatographic equipment. Alas no currently available chromatography is capable of completely resolving the hundreds of thousands of peptides generated by the digestion of a eukaryotic cell proteome.

Beside improvements in the performances of both MS and on-line chromatography, one logical way to increase the number of peptides successfully sequenced, is to first divide the protein population into multiple fractions. Such a separation can be carried out both before and after protease digestion, i.e. both at protein and peptide level, taking advantage of different physical and chemical properties of the analytes.

SDS-PAGE separates polypeptides according to their electrophoretic mobility, mainly determined by their length, and allows in-gel protease digestion of the resolved proteins. Isoelectric focusing (IEF) takes advantage of an immobilized pH-gradient to resolve polypeptides according to their overall pI (isoelectric point). This two methods were combined in 2D gel-electrophoresis (O'Farrell 1975; Görg et al. 2004) to obtain an orthogonal fractionation that has been the powerhouse of proteomic in its early history. Unfortunately, 2D-gel electrophoresis presents a strong bias against proteins localized in membranes, those with extreme pIs or molecular weights, and is associated with a very laborious workflow.

To overcome the shortcomings of 2D-gel electrophoresis, the group led by John Yates proposed a gel-free peptide multidimensional fractionation strategy based on orthogonal chromatographic separations (Link et al. 1999; Washburn et al. 2001). This setup couples reverse phase peptide resolution

with *Strong Cation exchange* (SCX), a chromatographic method based on the affinity of peptides for a negatively charged stationary phase. In principle the concept can be extended implementing any flavour of liquid chromatography (LC) to resolve the analytes according, for example, to their hydrophobicity (reverse phase/RP-LC), charge (SCX or SAX), or PTMs and ligands (affinity LC). Since ESI sources are sensitive to salts, RP-LC is often used in line with the mass spectrometer. Sample fractionation can also be driven by biological considerations as in the case of subcellular organelles enrichment (Brunet et al. 2003) or specific-cell sorting.

PAPER I presents the comparison of several proteome fractionation methods, both gel-based and gel-free, orthogonal to a reverse phase chromatographic separation. Two techniques were used which fractionated the samples after tryptic digestion, SCX and IEF, while intact proteins resolution was obtained by either SDS-PAGE or sub-cellular organelle separation. SCX has been used for many years as a first dimension of fractionation and additionally it can find application in sample clean-up and in PTMs enrichment (Beausoleil et al. 2004). IEF (Hörth et al. 2006) essentially applies at peptide level the same pI-based separation exploited in the first dimension of 2D-PAGE. Unlike SCX, IEF it does not require, nor tolerate, high salt concentration, which may interfere with peptides ionization in the MS. Beside intact proteins fractionation, SDS-PAGE coupled with in-gel proteolytic digestion removes from the sample many contaminants that might negatively affect the subsequent RP chromatography and MS analysis. In the same paper, in-gel digestion by two different proteases, trypsin and GluC, was also compared. Furthermore, as a biology-driven fractionation method, sub-cellular organelle fractionation by differential density centrifugation was adopted. This strategy combines increased proteome coverage with biologically relevant information on the sub-cellular localization of the identified proteins. Overall, SDS-PAGE followed by trypsin digestion yielded the best performance in terms of peptides and proteins identified. This result can be accounted for by four factors: i) resolving the sample at protein level confines very abundant peptides in fewer fractions; ii) SDS-PAGE is compatible with high detergent concentration, which in turns enhances sample solubilisation; iii) it effectively reduces contaminants, hence improving subsequent RP and MS analysis and iv) trypsin digestion produces peptides with better MS properties compared to GluC.

Of course the fractionation technologies are not restricted to those listed here. In theory any difference in the physical or chemical properties of the molecules being analysed can be exploited to divide them and/or, as it will be

discussed later, enrich for specific types of analytes. Antibodies in particular might provide yet additional means to resolve the sample. Besides binding entire proteins (see for an example: Lacey et al. 2001), antibodies can be used to capture specific peptides before their detection via MS (Scrivener et al. 2003). This idea found applications mostly in targeted proteomics, to enhance the quantitation of specific peptides via SRM (N. L. Anderson et al. 2004), and also on a MALDI platform (Jiang et al. 2007). However, variable specificity and sensitivity of antibodies turns their use to fish target peptides into a risky game: MS detection may indeed overcome many specificity-related issues but not the problems arising from the antibody not binding its target, which then results in false negatives. On the other side, motif-specific rather than sequence-specific antibodies show the potential to increase the output of shotgun proteomics providing an novel tool for sample pre-fractionation (Olsson et al. 2011).

## Quantitative shotgun proteomics

The huge dynamic range of protein concentration in cells and extra-cellular fluids reflects the fact that the *amount* of a given specie within a sample is critical to assess its functional significance. It follows that if the goal of proteomics is to characterize a protein sample with a global approach, the list of identifications should then be annotated with quantitative data. In most cases, the term "quantitation" is used to describe the relative comparison of protein levels between two or more sample (for example upon treatment). Absolute quantitation relies on the use of isotopically labelled standards with known concentration (Gerber et al. 2003; Havlis & Shevchenko 2004) to infer the protein amount defined as copies/cell or molarity.

As mentioned before, the TSQ mass spectrometer employed in targeted proteomics offers the best quantitation performances in terms of sensitivity, dynamic range and sample-to-sample reproducibility. This method of course is not suitable in discovery phase, i.e. to detect unknown proteins being differentially expressed, but is a valuable tool for subsequent validation.

Several workflows have been developed to retrieve quantitative data from shotgun proteomics: the most relevant are described here.

## Gel-based quantitation

2D-gel electrophoresis provided a first way to compare different samples by measuring the relative colorimetric intensity of stained protein spots. This procedure is limited by the high technical variability of the spot maps and by the relatively poor resolving power. Some of these shortcoming were alleviated by the introduction of DIGE (Unlü et al. 1997; Marouga et al. 2005), a technique based on protein pre-labelling by fluorescent cyanide dyes known as Cy2, Cy3 and Cy5. After labelling, the different samples are mixed and resolved on the same gel, therefore solving the issues connected with inter-gel technical variability. Identical proteins from different samples co-migrate in the 2D-gel but remain distinguishable using fluorescent imaging and the variation in their fluorescent read-out provides a measure of their relative concentration. Moreover, DIGE offers the possibility to align multiple gels by running in all of them a CyDye-labelled standard. The method has a good dynamic range and higher sensitivity compared to traditional visualization methods such as silver staining.

Unfortunately, while DIGE overcomes some of the reproducibility-related issues this approach still suffers of many of the 2D-gel electrophoresis drawbacks: the relatively low number of proteins visualized (and quantified), the tedious process of MS analysis of the spots, the bias against certain class of proteins and the ambiguous results obtained for co-migrating proteins. In PAPER II, DIGE is applied to relative quantitation of soluble proteome changes induced by hypoxic condition.

## Label-free quantitation

A major alternative to gel-based quantitation consists of using the mass spectrometer itself to infer the protein/peptide abundance. An important consideration to keep in mind though, is that the ionization properties of each peptide depends on its amino acids composition, which means that the ion current associated with any given peptide cannot be used as a direct measurement of its concentration. MS-based quantitation strategies can either be label-free or take advantage of stable isotope labelling. In the label-free approach, the samples are independently analysed using an identical workflow, and individual peptide properties are then compared. Some groups proposed to use secondary data from database searching, such as the protein identification score (Allet et al. 2004) or the number of peptides used to identify a certain

protein (Ishihama et al. 2005) or the frequency at which a peptide was randomly selected for MS/MS (H. Liu et al. 2004) as quantitation tool. Currently, methods based on feature intensity are gaining popularity in the label-free field mainly because of the higher accuracy achievable. "Intensity" here usually refers to the extracted ion current (XIC, i.e. the integrated peak area obtained plotting the ion current generated by a peptide over time) and for identical peptides and in the same experimental conditions is proportional to the ion amount. This approach requires a very reproducible sample preparation and resolution, and data analysis is usually complex and potentially prone to artefacts that may critically affect the outcome.

## Stable isotope labelling

The liaison between quantitative proteomics and stable isotope labelling dates much earlier than label free methods. Stable isotopes (typically $^{18}O$, $^{15}N$, $^{13}C$) do not affect chromatographic separation and ionization properties of the peptide they are incorporated in, but they introduce a MS-detectable *m/z* shift that can be used to label analytes from different samples, pool them and measure their relative intensity in the same analysis. Synthetic peptides containing uncommon isotopes can be added to the sample at known concentration to measure the absolute concentration of their natural counterpart (Gerber et al. 2003), but their use is practically limited by economical considerations to relatively small proteomes. When labelling of the whole proteome is required, the choice is between enzymatic incorporation during protein digestion, chemical attachment of an isotopic tag or metabolic uptake of the "heavy" isotope.

An example of enzymatic labelling was introduced by the Roepstroff's group: if the reaction is carried out in $H_2{}^{18}O$ water, the hydrolytic cleavage catalysed by trypsin can introduce $^{18}O$ at the peptides C-terminal (Mirgorodskaya et al. 2000).

The second way to achieve isotopic labelling of a proteome is to derivatise reactive groups on the protein structure with isotope-labelled reagents. A broad range of isotopic tags has been developed: here only few remarkable examples will be discussed. ICAT (Gygi et al. 1999) was the first reagent of this class to become available. In its first version it contained a cysteine reactive group, a poly-deuterated linker and a biotin group exploitable for peptide affinity enrichment. The selectivity for cysteine, a relatively rare amino acid, leads to a desirable sample simplification but at same time cause

many cysteine lacking proteins to remain undetected and often offers quantitation relying on a single or few peptides. Since trypsin is the most widely used protease in bottom-up proteomic, the labelling of primary amines (lysines and N-termini) would in principle affect all peptides. In PAPER II we use N-terminal and lysine labelling with either $^{2}H_4$ or $^{1}H_4$ -nicotinic acid derivative (D4/D0-Nicotinoyl-N-hydroxysuccinimide) for gel-free membrane proteins quantitation. The same N-terminal and lysine directed chemistry is used by iTRAQ (P. L. Ross et al. 2004) and TMT (Thompson et al. 2003), which also introduces a tag-specific reporter ion which gets cleaved during peptide fragmentation allowing quantitation from the MS/MS spectra. A variable linker between the reporter ion and the amine-reactive group ensures all different tags to be isobaric, hence ensuring co-fragmentation of identical peptides with different labels. The method has been further developed to include currently 8 different tags, thereby allowing further multiplexing.

## SILAC

PAPER III and IV exploit *Stable Isotope metabolic Labelling of Cell lines*, a technique made popular by Mann's group under the name of SILAC (Ong et al. 2002). Indeed, the idea of feeding "heavy"-isotope containing nutrients to micro-organisms grown *in vitro* in order to achieve whole cell labelling had already been used (Oda et al. 1999) but during the last ten years the protocols have been adapted to eukaryotic cells (Ong & Mann 2007) and reliable software tools have been made publicly available to readily analyse the data (Jürgen Cox & Mann 2008; Jürgen Cox et al. 2009). The basic concept is extremely straightforward: stable isotopes are supplied to proliferating (or anyway carrying on protein synthesis) cells replacing one or more essential amino acids with its heavy counterpart; the stable isotope should be chemically equivalent for the organism under investigation, so that no measurable change in protein expression should be induced; once the whole proteome is (nearly) completely labelled, different samples can be pooled already at a intact cell stage, virtually eliminating any bias introduced by subsequent steps of protein extraction and fractionation. However light and heavy peptides are detected by the mass spectrometer as distinct chromatographic peaks whose XICs are then used for relative quantitation (Fig. 2).

Any essential amino acid can in principle be the target for labelling but lysine and Arginine are most commonly used as they guarantee that virtually all tryptic peptides, except the protein C-terminus, will be labelled. Specifically: $^{15}N_4$- , $^{13}C_6$- and $^{15}N_4{}^{13}C_6$-L-arginine introduce a mass shift (referred to as "Δ" from now on in the following discussion) of 4, 6 and 10Da respectively. $^{2}H_4$-, $^{13}C_6$- and $^{15}N_2/^{13}C_6$-L-lysine are 4, 6 and 8Da heavier than the standard $^{1}H/^{12}C/^{14}N$ amino acid. Keeping in mind that at least a 4Da difference between the differentially labelled peptide are required to achieve a clear separation of the isotopic clusters, the combination of KΔ0/RΔ0, KΔ+4/RΔ+6 and KΔ+8/RΔ+10 gives the highest sample multiplicity of three, even though in duplex experiment deuterated lysine are generally avoided to avoid any peak shifting in RP chromatography.

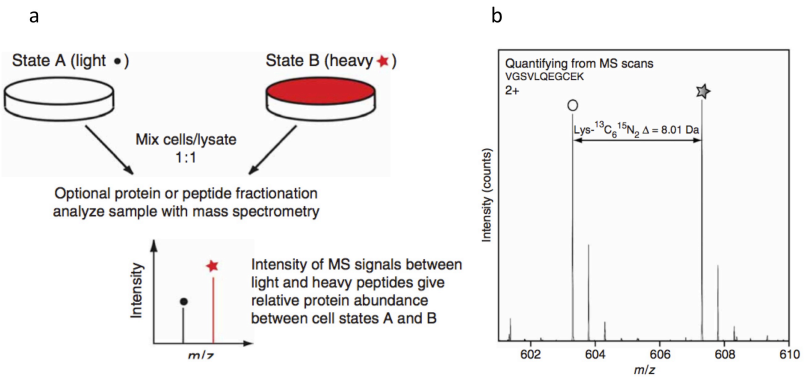


**Figure 2**
*SILAC rationale (a) and the m/z shift produced (b) (Ong & Mann 2007)*

Since it is independent of reaction kinetics (unlike enzymatic and chemical methods), SILAC produces very efficient and controllable labelling. The possibility to combine the samples at a very early stage allows extensive and unbiased fractionation strategies, which in turn leads to a better proteome coverage and higher number of quantitated peptides. Arginine-Proline conversion is a factor that could reduce peptide quantitation but methods have been developed to control this source of error (Van Hoof et al. 2007). The limited multiplexing capability and the limitation to cells grown *in vitro* have also been overcome with the introduction of the Super-SILAC idea (Geiger, Juergen Cox, Ostasiewicz, et al. 2010b). Moreover, this quantitation method

has been proven to be very flexible, its applications ranging from intra-cellular protein trafficking (Boisvert et al. 2010) to quantification of PTMs (Hilger et al. 2009; Ong et al. 2004). Even if such an experiment has obvious burdens of an economical nature, metabolic labelling of whole multi-cellular organisms such as *C.elegans* and *D. melanogaster* (Krijgsveld et al. 2003), plants (Engelsberger et al. 2006) and mammals (Wu et al. 2004; Geiger et al. 2013) has also been reported and opens new experimental possibilities.

## Difficult sub-proteomes

Despite the technological progresses presented so far, there are many classes of protein that remain difficult to be disclosed by standard proteomic techniques. Yet, sometimes these tenacious sub-proteomes hide highly relevant proteins that justify the extra-effort required to unveil them.

Two examples of such classes of proteins and of the methods available to survey them are discussed hereunder.

### Phosphoproteomics

Post-translational modifications (PTMs) are chemical changes to the structure of amino acids occurring in a protein after its ribosomal synthesis and play a central role in functional regulation. Most PTMs consist of the addition of a chemical group (for example phosphorylation, acetylation, methylation, sumoylation and so on) or in the removal of part of an amino acids side chain (for example the conversion of Arginine into citrulline) (Wilkins et al. 1999; Creasy & Cottrell 2004). Phosphorylation in particular is of paramount importance in controlling virtually all cellular processes, justifying considerable efforts to describe its dynamics on a whole proteome scale (Ubersax & Ferrell 2007). From a mass spectrometrist's point of view, the transfer of a phosphate group from ATP to a serine, threonine or tyrosine results in a specific 80Da increase in the modified residue mass. This shift can in principle be the diagnostic evidence driving the modification detection and the subsequent quantification by mass spectrometry (Zhao & Jensen 2009).

However, the measurement of phospho-peptides poses quite a few challenges. The first problem to be addressed concerns under-sampling of

phospho-peptides with low abundance and/or low site occupancy. As discussed above, fractionation and/or enrichment steps after the protease digestion of the sample are common ways to make peptides accessible to MS analysis. Leaving aside antibody-based capture, phosphate electronegativity is the feature most often targeted for enrichment, for example by SCX chromatography (Beausoleil et al. 2004) and *Immobilized Metal Affinity Chromatography* (IMAC, Andersson & Porath 1986; Neville et al. 1997), both taking advantage of the affinity of negatively charged phosphate groups (but also carboxylate and so on) for cations such as $Fe^{3+}$, $Zn^{2+}$, $Ga^{2+}$. Titanium dioxide (Pinkse et al. 2004; Aryal & A. R. S. Ross 2010) entered the stage more recently, promising a better performance in comparison to IMAC and is the method chosen in PAPER IV to investigate irradiation-induced variations in the phosphoproteome. A combination of multiple enrichment steps, for example SCX followed by $TiO_2$ (Olsen et al. 2006; Hilger et al. 2009), is likely to enhance the specificity but at the same time requires more starting material.

Enrichment is not the only nor the biggest challenge posed by phospho-peptides. The phosphate group itself often confers a poor MS behaviour to the modified peptides. Firstly, since mass-spectrometers are usually operated in positive mode (i.e., the source of the instrument is set to ionize the molecules into cations) the negative charge of phosphate prevents the efficient ionization of the peptide carrying it. Therefore, if no enrichment is preliminary conducted the unmodified peptide is often favoured for detection. Furthermore, even when a phospho-peptide is enriched enough to be selected for MS/MS, the neutral loss of phosphoric acid is often the most prominent signal recorded after CID (Syka et al. 2004). Insufficient fragmentation translates into poor, if any, peptide identification and does not permit the correct localization of the phospho-site. Keeping the standard hardware, Gygi and co-workers (Beausoleil et al. 2004) described a multistage activation with the detection of neutral loss at MS/MS triggering a further fragmentation step ($MS^3$) which drastically increased the number of observed daughter ions. Alternatively, improved structural characterization has been obtained changing the fragmentation method to ETD (Chi et al. 2007) or HCD (Olsen et al. 2007).

A further, even subtler, obstacle has yet to be overcome to quantitate phosphorylated proteins and assess for each PTM site the relative occupancy on a global scale: comparing different samples both the protein amount and the site occupancy can independently change. To give an example; comparing two samples A and B and pretending to have a perfect enrichment and detection. A given non degenerate phospho-peptide has in sample *A*, an XIC ten times

bigger than in sample *B*: how to interpret this result? The peptide can be 10 times more phosphorylated in A then in B (i.e. the site occupancy changed), or the protein that originated that peptide had in A a ten-fold up-regulation (i.e. the protein concentration changed) or a combination of the two. In principle only the quantitation of the non-modified peptides from the same protein could address this ambiguity but this information is often missing. On the same line of thinking, this last consideration has implication on quantitative proteomics outside the frame of PTM directed experiments: a differential concentration for a peptide may be caused by either a true variation in protein expression or by the occurrence of an undetected PTM.

PAPER IV describes a survey of phospho-peptides based on $TiO_2$ affinity capture. The enrichment strategy allowed the detection of many known phospho-sites as well as the identification of putative novel one. In this preliminary study, a large number of peptides were measured in the $TiO_2$-enriched fraction but for many of them no diagnostic peak in the tandem spectra was recorded. This on one side highlights the need for high-throughput strategy to confirm localization and extent of the modification, on the other hints at a lack of specificity of the $TiO_2$-enrichment strategy, which could possibly be addressed by pre-fractionation of the sample.

## Membrane proteomics

Gene sequence analysis revealed that about 40% of mammalian genes encode products with trans-membrane domain(s). These proteins comprise species likely to be exposed to the outer side of the cell membrane, thereby becoming obvious candidate as drug targets and biomarkers (Polanski & N. L. Anderson 2007).

As in the case of phosphorylated proteins, in membrane proteomics high biological interest comes together with consistent experimental challenges. The main issue in this case is represented by the amphipathic nature of membrane proteins, whose primary structure typically present a combination of water-soluble domains, often bearing PTMs, and highly hydrophobic trans-membrane stretches. The discussion about phospho-proteomics outlined some of the difficulties inherent PTMs analysis, which in the case of membrane proteins is complicated by the exceptionally complex modifications array observed in this class of proteins. Moreover, membrane proteins tend to be present at low abundance, further complicating their detection (Speers & Wu 2007).

Despite being located in the lipid bilayer, proteins structured as β-barrel (like *porins*) usually present a distribution of hydrophobic residues similar to that of soluble proteins. Unfortunately this is not the case for the majority of membrane proteins, which display highly hydrophobic transmembrane domains in α-helical conformation. The solubilisation of these proteins presenting both extremely hydrophobic and hydrophilic stretches of the primary structure provides a conundrum, as organic phases cause the water-soluble domains to drive protein precipitation while aqueous solvents are not suitable for the trans-membrane domain. At the same time the solubilisation phase must be compatible with down-stream analysis and high concentrations of detergents and/or organic solvents are not compatible with many fractionation methods and common protease. As an example, since strong detergents are not compatible with isoelectric focusing, the precipitation of membrane proteins remains one of the bottlenecks of 2D-gel electrophoresis. SDS-PAGE with in-gel digestion partly circumvents this issue by employing relatively high detergent concentration but allowing its later removal. In PAPER I in fact, peptides collected from 1D-gel fractionations show the highest retention time in reverse-phase chromatography, an evidence that markedly hydrophobic species were indeed successfully solubilised. Another strategy to circumvent the precipitation issue consists in separating the trans-membrane domains from the water-soluble ones. This goal have been achieved for example by proteolytically "shaving" the protruding parts of the proteins (Rodríguez-Ortega et al. 2006). Alternatively, proteolysis can be performed in denaturing conditions using cyanogen bromide (Kaiser & Metzka 1999) or proteinaseK (Hilz et al. 1975). Trypsin-independent digestion is attractive even for a second reason: trans-membrane domains are usually poor in charged residues, including Arginine and Lysine, and as a consequence upon tryptic digestion produce long peptides with poor MS properties. Unfortunately, CID ionization and fragmentation of peptides with no terminal lysine or Arginine tends to be weak and uninformative because of the lack of positive charges. In fact, cyanogen bromide was used by Washburn and co-workers in one of the first attempts to sequence membrane proteins with a gel-free proteomic approach (Washburn et al. 2001) but very few peptides from trans-membrane domains were indeed detected. However, chemical addition of a cationic groups at the N-terminus of non-tryptic peptides has later been proven to improve CID fragmentation of proteinaseK-digested membrane proteins (Jansson et al. 2008).

The combination of poor MS properties, low abundance and poor solubility makes the detection of membrane proteins more prone to suffer

under-sampling compared to water-soluble proteins. As in the case of phospho-proteomics, under-sampling of specific classes of proteins can be alleviated by enrichment strategies. Organelle separation, discussed in PAPER I, could in principle offer a mean to increase the relative amount of membrane fraction in the sample, for example by depleting cytosolic proteins. Rahbar and co-workers used a method employing cationic silica to capture the plasma membrane fraction before MS analysis (Rahbar & Fenselau 2004). A further method for membrane enrichment is based on the different affinity of this fraction for different water-soluble polymers at moderately high concentrations, such as PEG and dextran. Once the polymers have separated in two different phases, plasma membranes are preferentially solubilised by the more hydrophobic PEG-containing top fraction. Optimizing the concentration of the polymers, in particular that of PEG, it is possible to selectively enrich for the membranes of specific sub-cellular compartments with slightly different characteristics. For example, refined two-phase partitioning protocols consent to enrich plasma membrane with higher yields compared to other organelles (Schindler & Nothwang 2006). Under-sampling of integral membrane proteins can be reduced also by removing polypeptides that are only loosely associated with the lipid bilayer and by releasing the content of the lumen membrane enclosed organelles. Both goals can be achieved by washing the membranes with sodium carbonate.

Enrichment of membrane by two-phase partitioning followed by integral protein enrichment with sodium carbonate wash and subsequent controlled proteinaseK digestion and N-terminal labelling has been exploited in PAPER II.

## Proteomics: a science in its teens?

The publication of the first draft of the human genome in 2001 (Venter et al. 2001; Lander et al. 2001; International Human Genome Sequencing Consortium 2004) has been the finish line of a multi-year effort and at the same time the starting blocks for a still on-going technology development process, which in little more than ten years made third-generation whole-genome sequencing a matter of days (L. Liu et al. 2012). This immense improvement in the sequencing performances depends even, of course, on the

chemical features of DNA: an homogenous polymer, composed of only four monomers, that within a cell undergoes modest changes in concentration over time. The predicted human 23,000 genes (circa) are expected to be turned into about five times as many transcripts, mainly because of mechanisms such as alternative splicing (Roberts & Smith 2002), and transcriptional regulation is responsible for significant differences in RNAs concentrations. However current technologies still permit to monitor RNA dynamics on a global scale (Schena et al. 1996).

*Proteomics* shares with *genomics* and *transcriptomics* a suffix that alludes to a holistic approach to the study of the protein complement, but is this branch of science as mature as its *–omics* cousins?

To answer this question one needs first to define what a protein is. Is it simply the translation of a nucleotide coding sequence into amino acids? If so, the quest consists in being able to detect a number of different analytes ranging from 23,000 to about 120,000. Proteomics is not quite there yet, but given the pace of technological development it is reasonable to foresee that the target will be reached in the near future simply by refining the existing fractionation methods and MS instruments. This will be a major accomplishment, considering that proteins sequence is constituted by over 20 different monomers, instead of just four, and that the dynamic range can be three to seven orders of magnitude larger than that of transcriptomics.

In the introduction to this chapter, the stress was on the functional aspect of proteins: the efforts to characterize the final effectors rather than the gene encoding them are justified by the fact that the activity of a protein depends, among other factors, on its concentration and on its chemical and structural modifications (not to mention the cellular localization). Current techniques to acquire quantitative proteomics data have been introduced earlier in this chapter. Overall, a cautious optimism can be expressed about the possibility in the near future to quantitate significative portions of a cell proteome. At the same time, software tools developed to perform functional analysis of transcriptomic data can be adapted to the output of a quantitative proteomics experiment, as shown in PAPER III. This task can also be performed by FEvER, a novel tool to interrogate quantitative proteomic dataset looking for differentially regulated pathways that is presented in PAPER V.

Shifting the focus on PTMs, from a chemical and functional point of view each specifically modified proteins or peptides should be considered as a distinct entity. Under this premise, the number of analytes that proteomics is called to identify and quantitate has to be counted in millions. State-of-the-art tools for global profiling of PTMs are simply not sufficient for the titanic task

required to reach whole proteome coverage. Remaining within the boundaries of shotgun proteomics, only a fraction of a human proteome is currently identifiable, partly because of MS under-sampling and partly due to software limitations (Jürgen Cox et al. 2008). In fact all standard algorithms for peptides identification require the user to pre-set which PTMs to include in the search, a list that for practical reasons rarely encompasses more than two or three modifications. Finally, pretending *ab absurdo* that PTMs global detectability would suddenly improve, none of the informatics tools today available would be able to integrate these kind of data into a functional analysis.

From this perspective the current proteomic tools are still far from reaching the goal of a "whole proteome *functional* characterization". More reasonably, proteomics is currently enabling the identification of the primary structure of a remarkable number of proteins, for which full-coverage identification and quantitation seems achievable. Mass spectrometry has a unique potential in addressing proteins PTMs but the current paradigm is probably not suitable for global scale characterization.

# CANCER

Cancer is the leading cause of death in the developed world and, due to growing life expectancy and to the adoption of "western" cancer-associated life-styles, its incidence is climbing even in developing countries (Jemal et al. 2011). The term, together with its synonym "malignant tumour", describes a disease caused by cells proliferating in an uncontrolled fashion, invading the nearby tissues and settling in distal sites of the body to create metastasis. Cell over-growth is associated with the progressive impairment or loss of the function of the tissues affected, which eventually results in death. About 80% of cancer-related deaths are caused by *carcinomas*, malignant tumours spawned from epithelial tissues, while in the remaining 20% the diseases affect blood-forming cells (*hematopoietic* cancers), non-epidermal mesenchymal cell types (*sarcomas*) and nervous system (*neuroectodermal* tumours) (Weinberg 2007). The diversity of tissues that can originate tumours mirrors a broad heterogeneity in cancer biology and clinical features, with different cancers subtypes possibly originating from the same tissue and presenting, in turn, a high degree of phenotypic variance even within the same tumour. Nonetheless, Hanahan and Weinberg elegantly pinpointed a set of common traits of cancers (Fig. 3) which also constitute the milestones along the path of transformation of normal into malignant cells (Hanahan & Weinberg 2000; Hanahan & Weinberg 2011). The order in which cells acquire the different cancer features and the molecular effectors involved can vary but in order to display its malignancy a tumour must be able to: invade nearby tissues and metastasize, induce angiogenesis (non haematopoietic cancer only), contain cells that can both evade apoptosis and limitlessly replicate, ignore exogenous anti-growth signals while becoming independent from external replication stimuli. Moreover, cancer cells are characterized by genetic instability, deregulation of energetic metabolism and the ability to trigger inflammation response and at the same time escape destruction by the immune system.

As hinted above, the process of transformation of normal tissues into malignant tumours is a multi-step progression driven by sub-populations of cells acquiring one trait at the time and then obtaining from it a selective

advantage. This has at least two obvious consequences in clinical practice. The first involves the time dimension of tumourigenesis: cells require a relatively long time to acquire their malignant potential (it is estimated that many cancers may need about two decades from the initial lesion to produce detectable symptoms). Some cancer features, for example genetic instability, further increase the pace at which the others are acquired. In fact, early detection is one of the key factors affecting the outcome of cancer treatments, as it often allows the treatment of the tumour before it has fully developed its malignant potential and its resistance mechanisms. The second general problem connected with the cancer heterogeneity consists in the presence at the same time point of different populations of cells with different molecular features and malignant potential. Detection and treatment of cancer cells is further complicated by the surrounding normal and pre-cancerous neighbours. The multifaceted nature of cancers has a positive side as well, as all the "hallmarks of cancer" listed above offer in principle a therapeutic target.

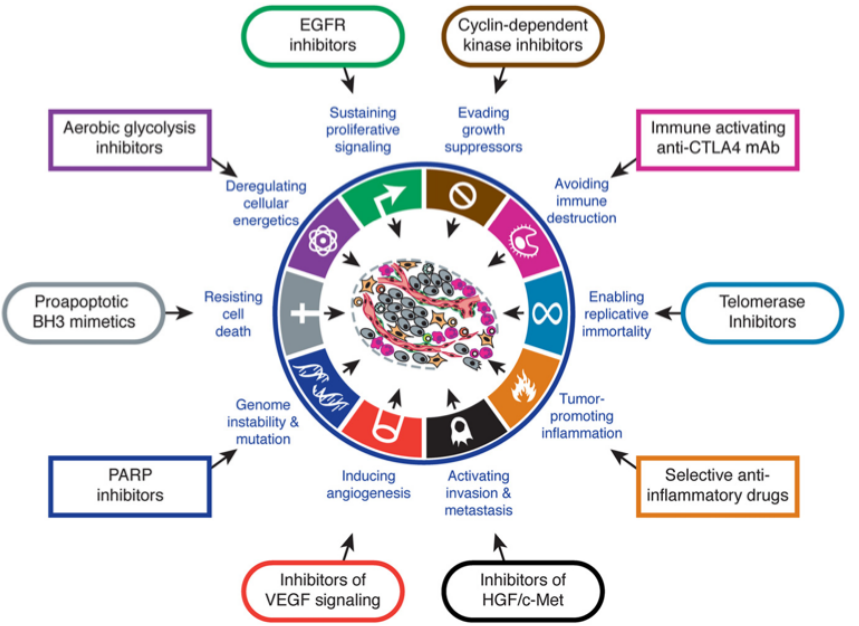


**Figure 3**
*The hallmarks of cancer (Hanahan & Weinberg 2011)*

Two of the hallmarks discussed above have been addressed in this thesis in relation to specific cancer types. PAPER II focuses on hypoxic conditions in neuroblastoma, while PAPERS IV and VI address the genome maintenance in breast cancer.

## DNA repair in breast cancer

In women, breast cancer is one the most frequently occurring and the second leading oncological death cause, after lung cancer (DeSantis et al. 2011). It is an extremely heterogeneous disease with distinct subtypes differing in the anatomical site of onset, molecular features and clinical outcome (Weigelt et al. 2010). It is not clear yet whether these differences define different diseases or separate steps of cancer progression (Simpson et al. 2005).

Twenty-odd breast carcinoma subtypes have been defined according to histological semblance and linked to different prognosis. Invasive *ductal* and *lobular* carcinomas account for 50-80% and 5-15% respectively of all diagnosed invasive tumours, and both classes also present sub-set of tumours with specific features. The remaining malignancies falls into sub-types with lower incidence (Weigelt et al. 2010). However, as of today there is still little consensus about the definition of such subtypes, and even the distinction between lobular and ductal cancer is being questioned (Simpson et al. 2005).

The main ranking scheme in clinical use is the *TNM* classification, which stages solid cancer according to *t*umour size, nearby *n*odes involvement and distant *m*etastatic status (Uehiro et al. 2013). Additional parameters, such as histological grade and hormone receptor expression, are normally included in this classification and TNM grades are often grouped into one of five stages ranging from 0 (carcinoma *in situ*) to stage IV (metastatic cancer).

Histological grading is based on morphological and cytological features that determine a combined score ranking the specimen as grade I (slow growing, well differentiated), grade II (moderately differentiated) or grade III (highly proliferating, poorly differentiated). The parameters taken into account, tubule formation plus nuclear pleomorphism for differentiation and mitotic index for cell proliferation (Elston & Ellis 1991), estimates the aggressiveness of the tumour and can therefore be considered a prognostic classification. Histological grading is affected by the subjective evaluation

operated by the human operator, an issue that could be alleviated by the definition of measurable molecular classifiers.

Expression of the oestrogen receptor (ER) was one of the first of such criteria for patient stratification to be defined and this feature was also found to be associated with major differences at transcriptomics level (Perou et al. 2000; van 't Veer et al. 2002; Sotiriou et al. 2003). At a later stage, Sørlie and co-workers proposed to classify breast cancers, according to the expression of a panel of "intrinsic" genes (Sørlie et al. 2001; Sorlie et al. 2003), into five subgroups: luminal A, luminal B, basal, HER+ ( or ERBB2+) and normal breast-like. Luminal tumours are the most common, express ER and have usually a relatively good prognosis. Luminal A and B cancers differ for the latter having a poorer clinical outcome, being less responsive to Tamoxifen (chemotherapy is in fact also recommended) and having a stronger proliferative signature. The HER+ subtype do not express hormones receptors and often display positive lymph nodes at diagnosis. Finally, basal-like tumours have the worst prognosis as they show very low or abolished expression of ER and HER, they express high proliferation genes and have usually high histological grade (Brenton et al. 2005). The *intrinsic gene* list was lately refined (Zhiyuan Hu et al. 2006) and yet additional subgroups have been proposed. Remarkably, how well the concentration of the transcripts in the intrinsic gene list correlates with the corresponding proteins levels has only marginally been assessed (Nielsen et al. 2004). A different scheme for prognostic classification was published by van't Veer and co-workers (van 't Veer et al. 2002) and this 70-genes marker panel is currently implemented in an FDA approved commercial assay (MammaPrint®) which estimates the risk that breast tumour will metastasize. A second molecular assay, Oncotype DX®, assesses the likelihood of recurrence according to the expression of 21 genes measured by RT-PCR (Paik et al. 2004).

Genomic aberrations have also been showed to correlate with clinical features and therefore proposed as a possible prognostic marker (Chin et al. 2006).

On the protein level, the concentrations of few proteins (ER, PgR and HER) are routinely used to stratify breast cancers in prognostic classes. In addition it has been reported that expression of cytokeratins 17 and 5 correlates with poor clinical outcome (van de Rijn et al. 2002). Expression levels of said cytokeratins, ER, HER and c-KIT consistently correlates as well with tumour defined as basal-like according to Sørlie's classification (Nielsen et al. 2004). The protein MUC1 has also been suggested as a prognostic marker for recurrence and responsiveness to chemotherapy (Duffy et al. 2010)

but its use in clinical practice is at present not recommended (Khatcheressian et al. 2013). Available biomarkers for breast cancers are also summarized in Figure 5 of the next chapter.

Despite the advances in cancer stratification, the treatment strategy is currently based on the relatively few elements included in TNM staging and on variables such as menopausal state, age and hormone receptor expression. It must be noted that slightly different treatment protocols are adopted in different countries and consequently this short overview will remain on rather general terms. Surgical removal of the tumour, by either lumpectomy or complete mastectomy, is basically the main treatment for non-metastatic cancers. Neo-adjuvant chemotherapy is often administered to reduce tumour size and facilitate its excision. Sentinel lymph node dissection and, if positive, axillary node dissection is also routinely practiced. Post-surgery systemic chemotherapy is generally administered to reduce the risk of recurrence, especially treating tumours with high Oncotype DX® score. Hormone receptors expression (10% or more of the cell staining positive for ER and PgR by IHC assay) drives the choice to give endocrine therapy, available as aromatase inhibitors and/or selective ER modulators like Tamoxifen. Trastuzamab is used in the treatment of HER over-expressing tumours. Radiotherapy, which is discussed more in-depth below, is also routinely delivered as post-surgery localized adjuvant treatment (Downs-Holmes & Silverman 2011; Yarnold 2009).

Many factors enhancing the risk to develop breast cancer have been listed, mainly related to events affecting hormonal status (for example parity and breast-feeding history, age of menarche and of menopause, use of oral contraceptives) and to environmental agents (ionizing radiation, fat-rich diet, alcohol consumption and so on, Dumitrescu & Cotarla 2005). In addition, family history is a well-known risk factor, since specific genetic markers are closely linked to the development of the disease. In particular mutations in BRCA (*Br*east *C*ancer *A*ssociated) 1 and 2 genes are often observed in families with cases of breast and ovarian cancer (King et al. 2003) while being relatively rare in sporadic cancers. Certain germ line mutations in TP53, PTEN, ATM, CHEK2 and LKB1 genes are well established cancer susceptibility markers, although with less penetrance than BRCA1 and 2 (Antoniou & Easton 2006). Remarkably, mutation in the BRCA genes and in TP53 are frequently found in basal-like aggressive tumours (Brenton et al. 2005). Most of these genes encode proteins involved in DNA damage response, directly linking cancer development to deficiencies in maintaining the integrity of the genetic information. Loss or malfunctioning of genome

caretakers leads to an increased probability of acquiring new mutations and, therefore, to gain even more of the hallmarks of cancer previously discussed. In consideration of the extremely low error rate of the DNA replication machinery, the majority of mutations can be ascribed to insults to the genetic material and to the following repair processes.

## DNA damage and repair

Cellular metabolism generates many chemical products that can react with the DNA molecules altering their structure and function. Moreover, DNA is exposed to exogenous threats ranging from radiation to genotoxic chemical species, whose damage adds up to the endogenous. The results of these insults include covalently bound bases, abasic sites, base adducts and breaks in one or both the phosphate backbones (De Bont 2004). Since damaged DNA is normally prevented from being replicated and transcribed, a number of mechanisms have evolved to repair the lesions while preventing the cell from progressing through the cell cycle.

At present, eight main pathways account for damaged DNA signalling and repair: Direct Damage Signalling (DDS), Direct Reversal Repair (DDR), Base-Excision Repair (BER), Nucleotide Excision Repair (NER), Mismatch Repair (MMR), Homologous Recombination repair (HRR), Non-Homologous End Joining (NHEJ) and Trans-Lesion Synthesis (TLS) (Ciccia & Elledge 2010; Milanowska et al. 2011). Most of these repair processes are carried on in a multi-step fashion by enzymatic complexes encompassing elements that i) recognize the damage and recruit the repair proteins, ii) reverse the lesion and iii) prevents the cell to progress in the cell cycle with unrepaired damages. Replication and transcription are often associated with detection of the damage, as stalled DNA/RNA polymerases are one of triggers of DNA repair response. On the other hand, DNA damage is tightly coupled with regulation of cell cycle progression. In healthy cells, detection of damaged DNA leads to stop in G1/early S phase to allow the repair to take place before replication. Above a critical damage threshold or upon un-repairable lesions, the cell usually activates the apoptotic program (Galluzzi et al. 2012), a process in which TP53 plays a pivotal role. It comes to no surprise then that many cancers exhibit a mutationally-inactivated TP53, which is no longer able to properly trigger DNA damage-induced apoptosis.

The accuracy of the repair process varies among the pathways above and indeed some of the mechanisms, such as NHEJ, are intrinsically mutagenic.

Some repair mechanisms are very lesion-specific (for example certain enzymes performing DRR), while others are partially redundant and can resolve a broader range of damage types. In many cancers in fact, the loss of function in one DNA repair pathway is often partially compensated by those remaining. These repair mechanisms are occasionally up-regulated providing the cancer cells with means to better resist genotoxic insults, including those received for therapeutic purposes. Thus in cells with impaired DNA maintenance, the effectiveness of genotoxic treatment might be enhanced by the simultaneous administration of drugs inhibiting some of the remained active pathways, a strategy named *synthetic lethality* (Bouwman & Jonkers 2012). Since normal cells are expected to display the full set of repair mechanisms, synthetic lethality should increase the sensitivity to genotoxic stress relatively more in tumour cells compared to the surrounding non-cancerous tissue. This approach is exemplified by the administration of PARP inhibitors to patients with mutated BRCA genes (McCabe et al. 2006).

## Radiotherapy and DNA double-strand breaks repair

So far the process of DNA damage and repair has been discussed mostly for its role in cancer onset and progression. The introduction of genotoxic therapies and of synthetic lethality hinted at the other side of the coin: the impaired ability of cancers to maintain their genomic integrity constitutes an Achilles heel as it may increase tumour sensitivity to genotoxic treatment (Jackson & Bartek 2009) in two main ways. First, actively proliferating cells, such as cancerous ones, undergo the DNA quality control checkpoints more often than their quiescent neighbours, increasing the likelihood of DNA damage-induced apoptosis. Secondly, inactivation of one or more DNA repair pathways translate into a reduced DNA repair capability, which means that healthy tissue around the tumour should be comparatively more resistant to the genotoxic stress. In clinical practice, two main types of cancer treatments act producing genotoxic stress: various types of chemotherapy (for example cysplatin, topoisomerase inhibitors, purines and pirimidines antimetabolites, DNA intercalating agents, (Downs-Holmes & Silverman 2011) and radiotherapy.

Two DNA-damaging chemotherapeutics, doxorubicin and methyl-methanesulphonate, have been used in PAPER VI to induce double strand breaks and DNA methylation respectively.

Radiotherapy (RT) consists of the delivery at the tumour site of ionizing radiation, X- or γ-rays, to produce localized cytotoxic damage. Radiotherapy at excision site is almost always offered after breast conserving surgery and is also delivered at the axillary region if sentinel lymph node biopsy is positive and axillary lymph nodes excision is impossible or not complete. The dosage is normally around 40-50Gy in total, delivered in daily repeated fractions of ca 2Gy over five weeks, followed in some protocols by a one-week boost dose. This treatment effectively improves the success of surgical excision of localized tumours and it even increases the survival rate of patients diagnosed with distant metastasis (Langlands et al. 2013). It is important to acknowledge that post-operative radiotherapy made lumpectomy as effective as total mastectomy in the management of early-stage breast cancers, allowing for less invasive treatment (Fisher et al. 2002). In nearly 100% of the cases side effects include fatigue and erythema. Telangectasia, arm lymphoedema, shoulder stiffness and cosmetics symptoms are more rarely observed. In addiction, radiations can negatively affect the underlying organs, such as heart and lungs, causing inflammatory manifestations and fibrosis which may arise even decades after treatment. Given the shortage of markers for radiation response, most of the patients undergoing cancer lumpectomy are currently given radiotherapy despite the fact that some of them do not achieve any beneficial effect from it (Makinde et al. 2012). HJURP (Holliday Junction Recognition Protein) and peroxiredoxin-I (plus related proteins) have been proposed as markers for radio-resistance (Zhi Hu et al. 2010; Woolston et al. 2011) but not yet clinically validated. Remarkably the clinical effectiveness of radiotherapy has also been directly put in relation to the tumour sub-type (Langlands et al. 2013) but it is not clear whether the phenotypes are due to the intrinsically higher proliferation rate of certain subtypes such as HER+ and basal tumours or on specific molecular mechanisms.

Ionizing radiation affects many cellular sites, such as membrane lipids and proteins involved in signalling pathways (Cohen-Jonathan et al. 1999) but the genetic material remains the most critical target. DNA damage induced by γ-radiation is caused either by direct ionization or via the generation of reactive hydroxyl radicals from water molecules next to the nucleic acid. Direct and indirect ionization together produces base damage, single and double strand breaks and DNA-proteins cross-links (Ward 1988). Out of these possible outcomes, base damage and to a lesser extent single strand breaks are the most frequent lesions but are also less toxic for the cell, as specific non mutagenic repair mechanisms (mainly related to nucleotide excision repair) can reverse them. Double strand breaks instead, despite occurring at a

relatively low rate, are the most genotoxic damage induced by radiotherapy (Kavanagh et al. 2013). The process of rejoining the resulting ends involves recombination events in two different pathways, depending on the presence of sister chromatids to be used as template.

Cells in G0/G1 phase can normally only exploit the non-homologous end-joining (NHEJ), which is intrinsically mutagenic (Shrivastav et al. 2008; Chapman et al. 2012). When a replicated chromatid is available instead to be used as template (i.e. normally in late S and in G2 phase of the cell cycle), homologous recombination (HR) provides a relatively accurate alternative to reverse double strand breaks and competes with NHEJ to accomplish the repair process. In the latter situation the choice of the repair pathways to be used depends on which complex binds first and stabilizes the strands ends, as well as on the presence of other accessory factors like 53BP1 (Brandsma & Gent 2012). The DNA ends produced by radiations-induced breaks can assume many configurations, often not directly ligatable. In this situation NHEJ offers better flexibility compared to HR (Y. Ma et al. 2005) turning the first pathway into a more versatile repair mechanism.

The general scheme for DNA repair introduced above is followed by both HR and NHEJ, which can then be used as an example (Fig.4). The repair process follows a spatially and temporally ordered scheme, with tight regulation of the effector proteins by PTMs (Misteli & Soutoglou 2009). Both pathways are initiated by sensor proteins, such as ATM and the MRN complex, which in turn recruit the machinery needed to carry on the repair process (Kavanagh et al. 2013). Shortly, during HR Replication Protein A binds the DNA free ends initiating strand invasion in the sister chromatid. Once strand invasion is stabilized, DNA polymerase δ extends the ends in a template-driven accurate process. Finally, the resulting Holliday junction is solved and DNA ligase eliminates the nick (Kavanagh et al. 2013).

In NHEJ, the DNA ends are recognized and bound by the Ku70/80 dimer, which in turn complexes and activates the serine kinase DNA-PK. Since radiation-induced double-strand breaks often result in non-ligatable ends, different proteins including various nucleases may intervene to degrade a small single-strand terminal sequence to expose regions of micro-homology that can then be used for subsequent annealing, gap-filling and ligation (Wang & Lees-Miller 2013). The mutagenic potential of this pathways resides both in the nucleasic activity needed to provide ligatable ends, which causes loss of short patches of DNA, and in the aspecific way adopted to anneal the available ends, that can result in chromosome aberrations and gene amplifications. Notably, a less well characterized *alternative*-NHEJ mechanism exists, which

involves PARP proteins, the MRE11 complex and ligases 1 or 2 (Simsek et al. 2011).

Damage detection by ATM/ATR also results in TP53 activation, which in turn halts the cell cycle to allow time for the repair to take place or induces apoptosis if the lesions are too severe to be reversed (Oberle & Blattner 2010). TP53 is particularly critical for transcriptional activation of p21 and consequently arrest in G1-phase, i.e. before DNA replication is initiated. A further DNA checkpoint at G2/M transition can cause the so-called mitotic catastrophe (Castedo et al. 2004) if unrepaired damages are still present.

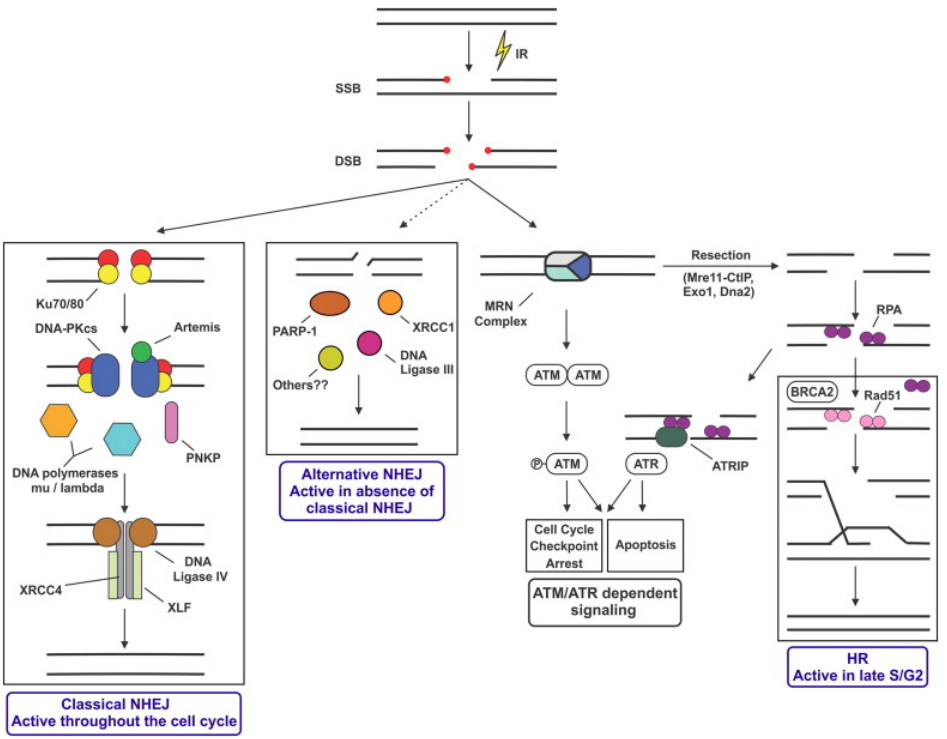


**Figure 4**
*Overview of the NHEJ and HR pathways (Wang & Lees-Miller 2013).*

At this point some molecular mechanisms possibly underlying genetic instability and apoptosis escape should start assuming a more definite shape. Mutations preventing TP53 from inducing p21 transcription prevent G1 arrest. Likewise, inactivation of ATM on one side reduces the efficiency of DNA damage sensing and repair but on the other prevents TP53 activation and the downstream cell cycle stop. Faulty BRCA1 or 2 results in defective HR and increase the mutation rate.

Molecular mechanisms responsible for radio-resistance occasionally observed in oncological practice involve more than just the DNA-repair process (Lewanski & Gullick 2001) but have not yet being completely elucidated. Tumour hypoxia, discussed later, is likely to play a role but a broader and clearer picture of the molecular survival mechanisms conferring tolerance to DNA damage is needed to improve treatment effectiveness.

PAPER IV quantitatively describes the proteome changes induced by γ-irradiations in two cell lines, one of which lacks functional TP53. The known response briefly outlined above was reconstructed based on protein levels and two further putative mechanisms of radio-resistance were proposed. First, the constitutive over expression of topoisomerases and, second, the activation of a DNA recombination pathway alternative respect to HR and NHEJ (SWAP complex) that is normally not involved in DNA repair. The activity of TP53 has been shown to clearly affect the response to the genotoxic stress but the cell line provided with the fully functional proteins was able to escape apoptotis as well, suggesting that additional cell cycle deregulations might be in place. To maximize the coverage of the DNA repair pathways affected by the genotoxic stress, a multistep sample fractionation based on i) cellular localization (nuclear enrichment), ii) protein size (SDS-PAGE) and iii) PTM status (phospho-peptide enrichment) was employed. Despite quantifying several thousands proteins only a partial coverage of the nuclear proteome was achieved and many expected phosphorylation sites were not observed. Another finding described in PAPER IV and VI is that one cell-line, MDA-MB-231 constitutively expresses DNA repair enzymes at higher level compared to MCF7. This observation has been obtained using different shotgun quantitation methods and confirmed by SRM in PAPER VI. Stimuli-independent activation of DNA repair pathways may then have a clinical significance contributing to confer resistance to genotoxic treatment.

In more general terms, mass-spectrometric detection of DNA repair proteins is useful not only in terms of increasing our understanding of cancer biology but also because it could have potential clinical applications even in the short term. The ability to detect at protein level deficiencies in one or more

DNA repair pathways (*BRCAness*) could be exploited to induce synthetic lethality in a patient-tailored fashion, one step closer to the goal of delivering a "personalized medicine". This task could practically be defined as the development of targeted proteomics assays able to assess presence and amount of relevant DNA repair enzyme, a goal that becomes a lot easier provided the corresponding tandem spectra are available. In fact, in PAPER VI a panel of SRM assays was employed to evaluate the proteomic response to two different chemotherapeutics. These assays can detect most of the DNA repair enzymes and cover all the pathways listed above. The experiments described in PAPER III (plus a preliminary study for PAPER IV) provided tandem spectra matching peptides from over 50% of the DNA repair enzymes listed in the RepairToire dedicated database, allowing to choose the best SRM transitions based on empirical fragmentation pattern.

## Hypoxia in solid tumours: neuroblastoma

Neuroblastomas are the most common childhood neoplasms, accounting for about 7 % of cancer cases and 15 % of related deaths in patients younger than 15 years old. This class of cancers originates from cells derived from the neuroectodermal lineage and the primary tumour can occur anywhere in the sympathetic nervous system. Symptoms produced by localized tumours are mostly connected to the physical compression of the nearby nerve but some cases are also associated with untractable diarrhoea due to the cancer secreting vasoactive intenstinal peptide (Maris et al. 2007). About half the patients are diagnosed with metastases in hematopoietic tissues. Amplification of the MYCN proto-oncogene is the most common genetic aberration observed in neuroblastomas (Puissant et al. 2013) and deletions in the short arm of chromosome 1 are observed in 20-25% of the cases (White et al. 1995) as well as other genomic aberrations.

### Tumour Hypoxia

Solid tumours, such as neuroblastomas but also breast carcinomas, often present poorly vascularised inner areas which experience low oxygenation (hypoxia) as well as nutrient starvation and low pH (Jögi et al. 2002; Brown & Giaccia 1998). This phenomenon is caused both by cancer cell growth

outpacing the development of blood vessels and by the erratic structure of the vascularization in cancer tissues. The distribution of oxygen within the tumour is therefore uneven, with only discrete areas suffering hypoxic stress. As a consequence, intra-tumour heterogeneity increases and clonal selection of cells with survival advantages is favoured, thus speeding up the gain of malignant traits (Axelson et al. 2005). Tumour hypoxia has clinical implications in terms of reduced sensitivity to genototoxic treatments and increased tendency to metastasize. The first phenomenon may depend both on insufficient drug delivery caused by poor vascularization and on the $G_1$/S arrest induced by the lack of oxygenation, which reduce the efficacy of therapies directed against proliferating cells. Moreover, quiescent hypoxic cells may form "dormant metastases" leading to relapsing of the cancer after treatment. Tumour hypoxia is also associated with radio-resistance, via the shortage of ionisable oxygen species enable to react with the DNA (Brown & Giaccia 1998). A further link exists between hypoxia and DNA damage: when new blood vessels are originated in the tumour, the shortage of oxygen is reversed, a phenomenon named *reoxygenation*, during which ROS are produced. Since cellular proliferation and angiogenesis proceeds at different paces within the tumour mass, repeated hypoxia and reoxygenation cycles keep producing ROS, which in turn cause DNA damage. As previously discussed, repairing lesions in the DNA is in part a mutagenic process itself. It follows that consecutive reoxygenation cycles may facilitate the acquisition of pro-tumorigenic traits (Vaupel & Mayer 2005).

The molecular mechanisms underlying tumour adaptation to hypoxic environment have been at least in part elucidated. *Hypoxia-inducible factors 1* and *2* (HIF-1 and HIF-2) play a pivotal role in sensing oxygen shortage and triggering the response. HIFs are constitutively expressed but, in normoxic condition, two proline residues in the $\alpha$-subunit are specifically hydroxylated by oxygen-dependent enzymes. The tumour suppressor VHL recognizes this modification, and in turn mediates the degradation of HIF$\alpha$. Molecular oxygen also inhibits binding of HIFs to their transcriptional co-activators. However, HIFs are stabilized under hypoxic conditions, when they exert their transcription activation function over hundreds of target genes, affecting processes such as: angiogenesis, cancer invasion, metastasis release, apoptosis and de-differentiation (Semenza 2003; Gordan & Simon 2007). Remarkably, in mice (but not *in vitro*) HIF and the related signalling are reported to be up-regulated even in oxygenated cells upon irradiation (Moeller & Dewhirst 2004).

A second major molecular mediator of hypoxic response is the *Nuclear Factor Kappa B* (NF-KB), which is induced both by low oxygen tension and by ROS. NF-KB activates pro-inflammatory cytokines and induces the over-expression of anti-apoptotic factors like BCL2 (Tamatani et al. 2000) contribute to confer to the cell a malignant phenotype.

Tumour hypoxia plays an important role also in breast cancer, which remains the main topic of this thesis. *In situ* breast carcinomas often presents anoxic/hypoxic necrotic areas surrounded by poorly oxygenated cell layers. This hypoxic tissue displays poor differentiation, previously discussed as a marker of poor prognosis, high HIF levels and down-regulation of oestrogen receptor in ER positive malignancies (Helczynska et al. 2003).

This brief overview offers a molecular link between hypoxia in solid tumours and their gain of malignant potential. This provided the rationale in PAPER II to look for hypoxic markers (biomarkers will be further discussed in the next chapter) in neuroblastoma-derived cell lines, as the proteome changes induced by poor oxygenation could constitute an alarm bell signalling the progression of tumourigenesis. The map of these changes obtained from other tumours, as in the case of neuroblastoma, might well help profiling the same response in breast cancer.

# CANCER PROTEOMICS

The development of *–omics* techniques has had a considerable impact on cancer research, resulting for example in the gene expression-based breast cancer classification previously discussed (Sørlie et al. 2001; van 't Veer et al. 2002).

The data presented in PAPER II and III shows that mRNA levels might not be of use to infer those of the corresponding proteins, which are needed to draw a phenotypic description. This means that transcriptomics data about cancer can still be useful for classification purposes but caution should be taken when describing cellular processes in terms of gene expression. Considering genomic markers, these features mostly define the risk or probability to develop cancer.

Can (shotgun) proteomics contribute to the field of cancer biology? And how? Proteins are the main class of molecules defining the actual phenotype (current status), so it is easy to answer a convinced "yes" to the first question. This chapter will discuss some aspects regarding "how" to conjugate proteomic and cancer research.

## Biomarkers

For many cancers, diagnosis of the disease relies on some sort of imaging technology (for example axial tomography, mammography, PET and pelvic ultrasounds) that is often employed after the first symptoms have already appeared, ruling out the possibility to deliver an early treatment. In the case of breast cancer, to be detectable by mammography a tumour need to be some millimetres in size, which correspond to millions of cancer cells. Moreover, these diagnostic procedures usually require invasive confirmation procedures (biopsies) and are not practical for population-scale preventive screening. One of the possible, and desirable, "outputs" of *omics*-sized studies would then be the definition of specific molecules being differentially expressed between two different biological states: shortly, the identification of "bio-markers" easily

and specifically detectable in body fluids such as blood (Taguchi & Hanash 2013) or nipple aspirate. Alternatively PET labelled antibodies against plasma membrane proteins could be developed. Cancer biomarkers (Fig. 5) might signal the mere presence of the disease (*diagnostic* markers) but also provide an outlook of tumour development (*prognostic* markers) or stratify patients according to other parameters such us drug responsiveness (*predictive* markers).

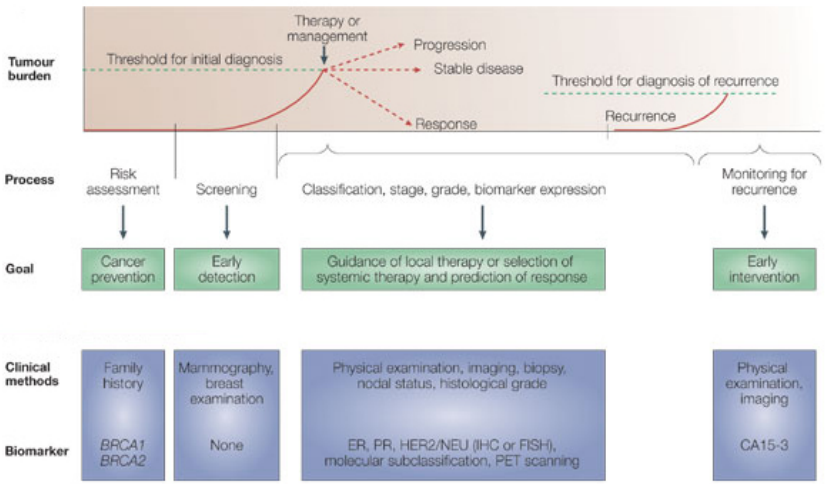


**Figure 5**
*Available breast cancer biomarkers (adapted from Ludwig & Weinstein 2005).*

The idea of fishing for cancer biomarkers using proteomic tools is basically as old as the word "proteomics" itself and the rationale behind is undoubtedly appealing and straightforward. Assuming the a cancer is in contact with the blood system and leaks macromolecules into it, the identification of tumour specific (or differentially expressed) proteins would provide biomarkers candidates to be further validated. Given the recent improvements introduced in the protein identification technologies one would expect the number of protein biomarkers identified to have proportionally grown. On the contrary, in parallel with the bloom of mass-spectrometry the rate of FDA approval of novel protein biomarkers remained steady averaging around 1.5/year (N. L. Anderson 2010) and only one accepted marker panel was specifically derived from proteomics (Zhang & Chan 2010). This

substantial "failure" has many reasons. The main one is related to technical limitations of mass-spectrometry based proteomics, shortly accounted for at the end of the relative chapter. The complexity and the huge range of protein expression levels in, for example, plasma samples (N. L. Anderson & N. G. Anderson 2002) still frustrates the quest for circulating biomarkers. Sample fractionation is the obvious way to go, but this plan collides with the need for an elevated number of samples required to obtain statistical significance and also with the limited amount of starting material that is generally available. Moreover, proteomic experiments aiming at biomarkers discovery often display inadequate experimental design and anyway few quantitative non-MS based assays are available for most of the human proteins to clinically validate the candidates molecules being proposed (Whiteaker et al. 2011). Mass-spectrometry based technologies have indeed the potential to overcome many of these shortcomings. Targeted proteomics could replace ELISA in markers validation (Addona et al. 2009) without the need for expensive *de novo* development of a test (Hüttenhain et al. 2012). At the same time, combining current mass-spectrometer with affinity enrichment (Razavi et al. 2012) shows the potential to detect peptides at the low end of the concentration dynamic range without time and sample consuming extensive fractionation.

## Towards a proteomic definition of "The Hallmarks of Cancer"

What has been said so far implies that the general strategy for biomarker discovery relies on the "blind" identification of proteins being differentially expressed in different clinically relevant states. In a sense this approach does not require any knowledge of, for example, cancer biology in order to define a biomarker. This strategy partly reflects the fact that at present even the most defined portraits of cancer are still rather blurry and contains many gaps.

Changing perspective, the complete molecular definition of the hallmarks of cancer discussed in the previous chapter would be the mother lode for rational biomarker discovery. The path to follow then would aim at defining cancer biology in proteomics terms, which means describing cancer processes at a global scale and possibly identifying a new generation of biomarkers in the framework of functional biology. Oversimplifying a little, one could say that the classical biomarker discovery workflow attempts to answer the question "Which proteins are differentially expressed?", a task that so far did

not prove fruitful. If instead the question would become "What are cancer cells doing differently?" (or, more properly, "which pathways are differentially regulated"), proteomics may return a more meaningful reply. A better and more integrate profiling of cancer pathways might then enable to obtain the overall picture from a relatively small set of measured data. The task that is being outlined here is indeed all but simple. Overlooking the obvious improvement in the number of detected analytes (i.e., the number of peptides, PTMs and so on), this approach requires integrating in the same global scheme quantitative proteomic data, PTMs detection and quantitation, sub-cellular protein localization and so on. This integrative, or "multi-dimensional" view calls for theoretical, informatics and analytical tools that are currently still to be developed. Most of the existing software in fact have been developed to suit transcriptomic datasets and therefore they do not make any use of data about detected PTMs or sub-cellular localization. What is more, these tools often adopt statistical analyses that do not fit proteomic results and do not directly handle entries from protein based databases. FEvER, the software described in PAPER V, can be considered as a first step towards the development of the informatics for proteomics-based functional studies. The program performs both an over-representation and an expression analysis on quantitative datasets, with the two strategies being combined in a flexible way to adapt to different data types. Information about protein PTMs and sub-cellular location are current not yet dealt with by FEvER, a feature that will hopefully be implemented in the future.

## What was sampled?

One of the main technical burdens in piecing together the functional map of cancer is the extreme complexity of this proteomic jigsaw puzzle. In clinical samples, both cancer heterogeneity and the number of cell types co-existing with the tumour (for example immune cells, tumour stroma and so on) contribute to further increase the overall complexity, resulting in multiple different proteomes being indeed analysed at the same time. This problem could be addressed by cell sorting technologies, or by adopting a model system constituted by cell lines grown *in vitro* (Neve et al. 2006). Both strategies can produce a homogeneous sample containing only one or very few cell types. However *in vitro* cultures also enable extensive fractionation strategies without

the limitation from the amount of starting material, metabolic labelling of the proteome (previously discussed) and a stricter control over environmental stimuli received by the cells.

In PAPER III the protein identifications obtained sampling breast cancer derived cell lines are compared to those from clinical biopsies analyzed with a similar experimental design. Only half the proteins of each dataset (therefore about one third of the total) were detected in both setups, a finding that was not really surprising. As anticipated, higher complexity of *in vivo* samples leads to more extensive under-sampling and, consequently, the mass-spectrometer misses low abundance proteins while is fragmenting peptides derived, for example, from the blood or from stromal tissue. Oppositely, focusing on cell lines very little information about "supra-cellular" and systemic processes was retrieved, but the functional picture of the cell processes was more detailed. This resulted in a finer dissection of cellular housekeeping processes, which remained largely unaccounted for when analysing tumour biopsies.

These results indicate that, keeping in mind the limitations still affecting state-of-art mass-spectrometry, cultivated cells may still contribute to the cancer proteomics field more than clinical samples by enabling to strictly control the stimuli received by the organisms under investigation. This idea is put in effect in three of the papers presented here. PAPER IV focuses on the cellular mechanisms leading to DNA damage tolerance. The novel mechanisms of radio-resistance there described (high constitutive expression of topoisomerase, presence of the SWAP recombination complex) constitute in fact a list of candidate markers with potential predictive usefulness to avoid the treatment of unresponsive patients. Genotoxic stress is also investigated in PAPER VI. PAPER II adds some details to the definition of the proteomic response to low oxygenation in solid tumours by focusing on the proteins localized on the plasma membrane and therefore most likely to be in contact with the outer side of the cell. This strategy led to the identification of several membrane proteins, including some of yet unknown function, to be further validated as hypoxic markers.

The differences between the cell proteomes *in vivo* and *in vitro* discussed with regards to PAPER III also reminds about some critical points connected with the use of immortalized cell lines as model system. The first matter of concern lies in the degree of similarity between cells *in vivo* and their cultivated counterparts. Many histological features of cell grown *in vitro* do resemble those of the cancers they derived from (Master 2000; Lacroix & Leclercq 2004), but it is hard to evaluate to what extent the two proteomes are

really similar on a global scale. Moreover, a single cell type culture cannot reproduce the complex relations occurring between the cancer cells and their environment (immune response, tumour stroma and so on). Co-culture of cancer and stromal cells (Miki et al. 2012) may alleviate this issue. PAPER III proves as well that most of the proteome of five different cell lines is indeed the same. This "core proteome" may either represent the "basic machinery" common to every cell, in which case its definition would be of high interest, or be the evidence of a common cellular response to the very same stimuli, meaning that experimental conditions masked the differences between the proteomes making them undetectable.

Overall, the limits connected with employing cultivated cells rather than clinical sample do not invalidate the evidences collected. However, extreme care should be used when transferring the knowledge acquired on cell lines straight into cancer biology.

## Predictably unpredictable

There is a slightly paradoxical note in trying to model malignant processes with a shotgun proteomic approach, as both the mission and the tools selected to accomplish this aim suffer from the somewhat anarchic traits of cancers. While tumorigenesis consists in acquiring a known and discrete set of "abilities" (Hanahan & Weinberg 2011), the order in which these traits are acquired is not predictable and the molecular mechanisms involved are not known yet. So, the problems posed by trying to achieve a "proteomic definition of 'The Hallmarks of Cancer'" are not merely of technical nature.

There remains a second paradox concerning in particular shotgun proteomics strategy relying on databases of theoretical spectra. As previously described, this workflow envisages to match the measured fragmentation with a theoretical spectral library created by *in silico* digesting and fragmenting a consensus version of the human genome. Alas, genomic instability and high mutation rate are the signature of cancer (Teschendorff & Caldas 2009; Weir et al. 2004). As a consequence, a point mutation changing even just one amino acid residue of a peptide is likely to prevent its identification. To give a practical example of the problem, in both PAPER III and IV the identification of TP53 is reported in MDA-MB-231, a cell line known to be carrying a mutated version of the *p53* gene; pretending that good tandem spectra had

been collected for every tryptic peptide generated from TP53, only those conforming to the standard *wild type* sequence could have been successfully and correctly identified, while those displaying the mutated sequence would not have been correctly assigned. The group of Richard Simpson recently addressed this problem by creating a database of recurrent mutations (Mathivanan et al. 2012). Most remarkably, they point out that cancer specific mutated proteins could also be useful biomarkers.

## Finding the way with a proteomic map

Most of the work presented in this thesis follows the scheme of a DDA shotgun proteomic with peptide identification based on database searching. At the same time these data may well merge into existing repositories (Vizcaíno et al. 2010; Mead et al. 2007) and/or become the core of a breast cancer-centric spectral library. This term has been briefly introduced in relation to peptides identification but it is worth understanding how spectral libraries can contribute to cancer proteomics.

The concept of mapping the entire proteome dates back over three decades ago, before the Human Genome Project had even been launched and with MS-based proteomics still out of sight, when Norman and Leigh Anderson proposed the Human Protein Index (N. G. Anderson & N. L. Anderson 1982; N. G. Anderson et al. 2001). While the original goals where mostly related to medical use, the later development of proteomics made the concept attracting even for more technical reasons (N. L. Anderson et al. 2009).

The importance of PTMs for functional understanding has been repeatedly stressed in the discussion so far and the large-scale identification of these modifications has been set as one of the main goals for proteomics. Alas, database searching is inherently ill-fitting to hunt multiple PTMs, as each variable modification considered exponentially expands the searching space (Lam 2011). Practically in order to retrieve peptide identifications with acceptable score only few modifications, normally two or three, can be set as variable in a database search. Vice versa, once the spectra of modified peptides are successfully assigned and stored in the library, this empirical database allows the identification of all PTMs at once while remaining significantly smaller than the corresponding theoretical one.

A second element limiting the process of matching empirical spectra with their predicted counterparts resides in the sometime loose correlation between the two. More precisely, a relevant fraction of the predicted daughter ions are often missing in the empirical spectrum because, for example, of poor ionization property of the fragment. On the other side, the "real" spectra contain unpredicted yet valuable peaks, corresponding to internal ions series, and so fort. In fact, relatively few peaks of each tandem spectrum usually support identifications by database search. Assuming that any given peptide would maintain its fragmentation pattern under identical experimental condition, spectral library search has the potential to improve the assignment of measured spectra, especially if of low quality or anyway "troublesome", respect to database searching.

Last but not least, spectral libraries provide evidence about the MS behaviour of the indexed peptides and therefore are greatly beneficial for the development of assays in targeted proteomics (Prakash et al. 2009; Hüttenhain et al. 2012). This task can in principle be accomplished even by using publicly available spectral repositories (Vizcaíno et al. 2010), such as *PeptideAtlas* (Deutsch et al. 2008; Picotti et al. 2008), *PRIDE* (Vizcaíno et al. 2013) or the *Global Proteome Machine DataBase* (Craig et al. 2004) but the development of an in-house library allows to customize the data annotation to fit the specific needs of each lab and ensure the experimental conditions to be more similar to those normally adopted.

These three elements would all enhance breast cancer proteomics. A higher efficiency of PTM detection would lead to a more meaningful functional profiling and is also likely to increase the fraction of tandem spectra successfully matched to a peptide. Peptide identification is also expected to benefit from "rescuing" spectra with difficult or poor fragmentation pattern, which in turn could increase the dynamic range of shotgun proteomics. Finally, targeted proteomics shows the robustness and reproducibility needed for biomarker validation plus it allows for fast fractionation-free workflows, making a bid for becoming the MS-method of choice in clinical practice. Mapping the cancer proteome would then provide tools to quickly translate basic research into clinical tools.

The potential of indexing the cancer proteome is not limited to enhance protein detectability. A wise annotation of the protein mapped could in fact narrow the gap between proteins characterization, in the broad definition adopted from the end of the chapter on Proteomics, and the assessment of their functional role.

# CONCLUSIONS AND FUTURE PERSPECTIVES

The papers on which this thesis is based outline some strategies to enhance resolution and detail of the breast cancer proteomic map.

PAPER I aimed at increasing the number of species selected for fragmentation in the MS, which led to a proportional raise in identified peptides. SDS-PAGE with in-gel tryptic proteolysis proved itself the most efficient out of the sample resolution methods tested. These results indicates how to relatively increase the number of identified peptides (and consequently proteins) regardless the LC-MS setup. The outcome of these analyses also directed the choice of the fractionation strategy adopted in PAPER III and IV.

PAPER II mapped the cancer proteome applying two different rationales: a biology-driven one and a second prompted by the need to overcome technological limitations. The latter consisted in targeting a specific sub-proteome that normally remains poorly characterized with standard proteomic tools: membrane proteins. Two-phase partitioning with subsequent proteinaseK digestion and labelling with NicNHS produced the identification of several membrane proteins, including some with no previous evidence at protein level. This workflow was employed to survey the proteome changes induced by tumour hypoxia, the biological rationale above. Of clinical relevance, PAPER II produced the identification of putative protein biomarkers for oxygen shortage in cancer cells and suggests a link between hypoxic stress and changes of expression of DNA repair enzymes.

In PAPER III the cells "standard" proteome was extensively sampled to retrieve redundant high-quality spectra. The described dataset portrayed a wide portion of the basic cellular machinery, not necessarily involved in malignant behaviours. Many of these proteins were identified with relatively high sequence coverage, providing empirical spectra for a portion of the proteome that is hardly reachable in clinical specimen due to sample complexity. These spectra can greatly facilitate the development of SRM assays for all the identified proteins, should any of them be measured *in vivo*, with potential usefulness for a wide range of applications.

PAPER IV follows the same two rationales described for PAPER II. The challenging sub-proteome analysed here was that undergoing phosphorylation. Modified peptides were enriched using their affinity for $TiO_2$ allowing the detection of known as well as novel phosphorylation sites involved in the response to γ-radiation (the "biological" stimulus). In particular, few DNA repair pathways were dissected that may be implicated in resistance to genotoxic therapies used in clinical practice. Quantitative data presented in PAPER IV needs to be validated in cell lines first and possibly in clinical samples too. The method to do so is already in place, with the SRM assays developed for PAPER VI.

PAPER V takes mapping one step further by trying to link the output of MS analysis with functional processes. Software for functional analysis may partly alleviate the shortcomings caused by MS under-sampling, by producing a global picture of the cell phenotype inferred from the relative few measured elements.

Finally, the potential of proteomic mapping is shown in PAPER VI, in which spectra collected in PAPER III (and in a preliminary experiment for PAPER IV) were used as a reference to quickly develop SRM assay to survey DNA-repair enzymes.

## Proteomics as a grown-up science: which way to go?

The technological platform of proteomics is still under development and as a result most of the studies published are still technology-driven, arguably reducing the output in terms of biological knowledge. The basic concept of this thesis consist in the idea that mapping the cancer proteome would allow to shift towards hypothesis-driven proteomics by enabling first to detect a larger part of the proteome and secondly to integrate the collected data producing biologically relevant information. However, both aspects need to be improved.

If the goal was to detect each *gene product*, then proteomics would have been very close to the finish line. However proteomics, and particularly MS-based proteomics, has an unequalled potential for global survey of PTMs and thus for producing functional description of samples. Counting PTMs bearing peptides as distinct entities though, a single eukaryotic cell potentially originate millions of different species to be detected, which means that with

current technologies a large part of the proteome still remains undetectable. Consequently, refinements of the methods for sample fractionation upfront the mass spectrometer and PTMs enrichment are needed to disclose yet new regions of the cancer geography. More generally, any technological advance alleviating MS under-sampling would have immediate positive consequences.

Data integration is the second field of improvement introduced above. Current tools for functional analysis define metabolic pathways in terms of proteins involved, which is a sensible approach if one could only detect the primary structure of proteins. However, this strategy "wastes" the potential of MS-proteomics. It is crucial to develop new ways to map the proteomic data and effectively integrate the different coordinates of the proteome complexity: protein primary sequence, concentration, PTMs, subcellular localization and so forth. Metaphorically, if we consider a cell like a machine it is surely important to identify where the toggles are (i.e. for example to detect phospho-sites) but we need to be able to profile the position of each switch over time and to link each on/off status with a different output.

Improving the specifications of MS-related technologies is a quest that necessarily requires a coordinated effort from the whole proteomics community. Data integration instead would benefit even from relatively inexpensive conceptual advancements and the works presented in this thesis can contribute to this development process. For example, the dataset presented in PAPER III needs to be indexed and made publicly available and at the same time may well be used to test novel more effective way to arrange proteomics outputs.

# POPULÄRVETENSKAPLIG SAMMANFATTNING

Uppsättningen av gener i DNA, *genomet*, kan betraktas som drivande för konstruktionen av varje levande varelse, medan motsvarande proteiner, *proteomet*, är materiellt ansvariga för de flesta biologiska processer. Kort sagt, gener bestämmer vad celler ”ska kunna göra” medan proteiner definierar vad varje cell "gör" ur en funktionell synvinkel. Aktiviteten hos dessa molekyler beror dock inte bara på deras sekvens, som bestäms av den gen som kodar för dem, men också på en rad faktorer som inte kan detekteras genom genetisk analys såsom koncentration, kemiska modifieringar och lokalisering i cellen. Undersökning av proteomet på global nivå kallas *proteomik* och syftar till att beskriva fysiologiska processer, framförallt patologiska sådana, genom att karakterisera de proteiner som är ansvariga snarare än de gener som kodar för dem.

Tack vare tekniska framsteg under de senaste åren är det nu möjligt att sekvensera hela genom på relativt kort tid. Samtidigt har utvecklingen av masspektrometri och relaterade tekniker gjort det möjligt att identifiera tusentals proteiner från extremt små provmängder. I den vanligaste versionen av denna teknik, digereras alla proteiner i provet till fragment, *peptider*, vars molekylvikt och fragmenteringsmönster mäts med hög precision. Dessa egenskaper tillåter identifiering av proteinerna och ger en uppskattning av deras koncentration för att sammanställa en proteomikkarta för funktionell förståelse. Hittills är karakterisering av hela det mänskliga proteomet dock fortfarande utom räckhåll på grund av den enorma komplexiteten hos de behandlade proverna och att den teknik som finns tillgänglig fortfarande är otillräcklig.

Ett av de områden som skulle gynnas mest av proteomik är forskningen om bröstcancer, en av de vanligaste cancerformerna bland kvinnor. Sjukdomen uppvisar en extremt heterogen patologi och är inte heller fullständigt karakteriserad ur en molekylär synvinkel. Dechiffrering av cancerproteomet är dett kritiskt steg för att definiera nya läkemedel, för att optimera de behandlingar som redan finns och för att identifiera markörer som skulle möjligöra tidig diagnos.

Den röda tråden som löper genom artiklarna i denna avhandling är sökandet efter strategier för att utöka den proteomiska kartan av just bröstcancer. Den första artikeln (PAPER I) presenterar en jämförande analys av fraktioneringstekniker av peptider och proteiner för att öka antalet identifierbara analyter per prov. I den andra artikeln (PAPER II) behandlas tumörcellers anpassning till förhållanden med lite eller inget syre genom att fokusera på en klass av proteiner som är svåra att analysera: membran proteiner. En kvantitativ karakterisering av de mekanismer som används av tumören för att vara motståndskraftig mot strålterapi (och till DNA-skadande ämnen i allmänhet) beskrivs i den fjärde artikeln (PAPER IV). Slutligen innehåller PAPER III grunden för sammanställandet av ett proteinindex för bröstcancer, ett verktyg som kommer att underlätta framtida analyser inom proteomik. Den sistnämnda artikeln, tillsammans med PAPER II, visar att analys av mRNA (de molekyler som fungerar som mellanhänder för kodningen av gener till proteiner) är otillförlitlig för att bestämma den intracellulära koncentrationen av motsvarande protein. PAPER V introducerar ett program för identifiering av de biologiska processer som pågår i cellen, grundat på data från proteomiska experiment. Slutligen beskriver PAPER VI tillämpningen av den "proteomiska bröstcancer kartan", som erhållits från tidigare artiklar, för att utveckla tester som snabbt, exakt och reproducerbart mäter specifika proteiner som ansvarar för DNA-reparation och potentiellt motstånd till radio- och kemoterapi.

Sammantaget tecknar artiklarna i denna avhandling konturerna av strategier för att göra cancerproteomet mer tillgängligt och mätbart. PAPER II och IV utökar dessutom förståelsen av två specifika aspekter av tumörbiologi.

# SUMMARY

The complete set of genes in the DNA, the so called *genome*, can be regarded as the blue-print directing the "building-up" of every living being, while the corresponding proteins, the *proteome*, are materially responsible of most biological processes. Shortly, proteins functionally define what each cell "does". Proteins activity depends not only from their amino acidic sequence, which is encoded by the relative genes, but also from several other elements that are not detectable by genetic analysis: concentration, post-translational chemical modifications, localization in the cell and so forth. The large-scale study of the proteome is named *proteomics* and aims at defining physiological processes, including diseases, by characterizing the proteins involved rather than the genes encoding them.

Technological advancements occurred in recent years have enabled sequencing entire genomes in relatively short time. At the same time, the development of protein mass-spectrometry and of the related techniques has allowed identifying thousands of proteins from very little amount of sample. The latter technology often consists first in the digestion of the proteins to be analysed in fragments, named *peptides*. Then, extremely accurate measurements of the peptides molecular mass and fragmentation scheme allow identifying the proteins that originated them and estimating their relative concentration. However, characterization of the entire human proteome is still out of reach because of the enormous complexity of the samples and of insufficient technical tools.

Proteomics techniques could be highly beneficial for the research on breast cancer, one of the most common malignancies in women. This cancer is an extremely heterogeneous disease still not fully characterized at molecular level. Deciphering the breast cancer proteome is a critical step to define new therapeutic targets, to optimize the delivery of already available treatments and to define new markers for early diagnosis of the tumour.

The *fil-rouge* linking the papers in this thesis is the search for strategies to enhance the resolution of the proteomic map of breast cancer. PAPER I presents a comparative analysis of some sample fractionation protocols commonly used to increase the number of proteins identifiable. In PAPER II a

specific biological process, the adaptation of tumour cells to oxygen shortage, has been addressed focusing on the a class of proteins generally difficult to analyse: the membrane proteins. Likewise, mechanisms adopted by cancer cells to resist radiotherapy (and other DNA-damaging agents) are investigated in PAPER VI. PAPER III aims at indexing breast cancer proteins to later build a dedicated database, a tool that will foster future proteomic research. This article, together with PAPER II, shows as well that measuring messenger-RNAs (the molecules that mediates the process of translating genes into proteins) is not a reliable method to determine the cellular concentration of the corresponding proteins. PAPER V introduces a software tool to infer the biological processes ongoing in the cell from the results of proteomic experiments. Finally, PAPER VI describes an application of the proteomic map of breast cancer drafted in the previous works: based on the data collected for the previous papers, accurate and reproducible assays for the measurements of proteins involved in DNA repair (and potentially in resisting many anti-cancer therapies) were developed. Overall, the papers presented in this thesis outline strategies to make cancer proteins easier to detect and quantitate and, at the same time, offer new insights to better understand some critical aspects of cancer biology.

# SINTESI
# A SCOPO DIVULGATIVO

L'insieme dei geni codificati dal DNA (il cosiddetto "genoma") può essere paragonato al progetto che guida la "costruzione" di ogni essere vivente, mentre le corrispondenti proteine (il "proteoma") sono materialmente responsabili della maggior parte dei processi biologici. In breve, sono le proteine che definiscono da un punto di vista funzionale quello che ogni cellula "fa". L'attività di queste molecole dipende tuttavia non solo dalla loro sequenza aminoacidica, determinata dal gene che le codifica, ma anche da una serie di fattori non rilevabili dall'analisi genetica quali concentrazione, modificazioni chimiche post-traduzione, localizzazione nella cellula e così via. Lo studio su scala globale del proteoma è detto "proteomica" e mira a descrivere i processi fisiologici, a partire da quelli patologici, caratterizzando le proteine che ne sono responsabili piuttosto che dei geni che le codificano.

Grazie ai progressi tecnologici degli ultimi anni é ormai possibile sequenziare interi genomi in tempi relativamente brevi. Allo stesso tempo, lo sviluppo della spettrometria di massa e delle tecniche ad essa collegate ha reso possibile l'identificazione di migliaia di proteine a partire da quantità estremamente ridotte di campione. Quest'ultima tecnologia, nella sua versione più diffusa, prevede di digerire le proteine del campione in frammenti, detti peptidi, di cui si misura con altissima precisione il peso molecolare e lo schema di frammentazione. Queste proprietà permettono di risalire all'identità delle proteine analizzate e, generalmente, forniscono un'indicazione sulla loro concentrazione. Ad oggi tuttavia, la caratterizzazione dell'intero proteoma umano continua a rimanere fuori portata a causa dell'enorme complessità dei campioni trattati e di ausili tecnici ancora inadeguati.

Uno dei settori che potrebbe beneficiare maggiormente dall'applicazione di tecniche di proteomica è la ricerca sul cancro al seno, una delle neoplasie più comuni tra le donne. Si tratta di una patologia estremamente eterogenea e non del tutto caratterizzata da un punto di vista molecolare. Decifrare il proteoma del cancro è un passaggio critico per definire nuovi bersagli terapeutici, per ottimizzare l'applicazione dei trattamenti già disponibili e per individuare marcatori che permettano la diagnosi precoce del tumore.

Il *fil-rouge* che attraversa gli articoli di questa tesi è proprio la ricerca di strategie volte ad aumentare la risoluzione della mappa proteomica del cancro al seno. Il primo articolo (PAPER I) presenta un'analisi comparativa di diverse tecniche per il frazionamento di campioni proteici, allo scopo di aumentare il numero di peptidi identificabili per campione. Nel secondo articolo (PAPER II) uno specifico problema biologico, l'adattamento delle cellule tumorali a condizioni di scarsa o nulla ossigenazione, è stato affrontato puntando l'obiettivo su una classe di proteine difficile da analizzare: le proteine di membrana. La caratterizzazione quantitativa dei meccanismi impiegati dal tumore per resistere alla radioterapia (e più in generale ad agenti dannosi per il DNA) è discussa nel quarto articolo (PAPER IV). Il PAPER III pone le basi per la costruzione di un "indice" delle proteine del cancro al seno, uno strumento che faciliterà future analisi proteomiche. Quest'ultimo articolo inoltre, insieme al PAPER II, dimostra che l'analisi degli RNA messaggeri (le molecole che fungono da intermediari nel processo di traduzione dei geni in proteine) è poco affidabile per determinare la concentrazione intracellulare delle corrispondenti proteine. Il PAPER V introduce un software per individuare i processi biologici in corso nella cellula partendo dal risultato di esperimenti di proteomica. Infine il PAPER VI descrive l'applicazione della "mappa proteomica del cancro al seno" ottenuta dai precedenti articoli: sfruttando i dati presentati negli altri articoli descritti, per il PAPER VI sono stati sviluppati dei test che misurano specifiche proteine, quelle responsabili della riparazione del DNA (e potenzialmente della resistenza a radio e chemioterapia), in modo rapido, preciso e riproducibile. Complessivamente, gli articoli presentati in questa tesi delineano strategie per rendere le proteine del cancro piú accessibili e misurabili e allo stesso tempo, aggiungono dettagli utili alla comprensione di due specifici aspetti della biologia del tumore.

# ACKNOWLEDGEMENTS

At some point of this thesis I defined proteomics a *science in its teens*. Likewise, these years as a PhD student had many traits of a scientific teenager-hood, with high expectations and stinging disappointments, plans to change the (proteomic) world and moments of deep depression. Some people accompanied me in this path and I want to use this space to thank them all.

First of all my gratitude goes to Peter, who let me take my time to fail and try again while always showing me I was entrusted ("I do believe your data"). Future will tell how well this nursing style worked.

Linn has been simply a lot more than a colleague. Without her I wouldn't have survived in both the Australian and the Swedish jungle. Peter enrolled her in a sociological experiment on cultural-stress tolerance and she performed amazingly well.

Every teenager gang has its geeks and mine had a very special couple of them. The chats I had with Ufuk left more than few traces in this thesis and helped shifting my perspective on many problems ("One recombinant frog, three legs,..."). Marianne, well, it has been a pleasure to host her pit-stops on the way to Fredrik and to share the same circadian time zone. Thanks for revising my Swedish.

Smiling at somebody who just killed the LC is commendable even if it happens just once. Karin and Mats did it with me, over and over and over again: thank you so much!

Sofia, our "almost grown-up" (no nicknames here): thanks for inspiration and for proofreading of the thesis.

Emila, the only person who would stop three buses packed with scientists on their way to dinner just to wait for the humble writer. We were already at the main course but thanks anyway.

Thanks to all the dark-side crew: Fredrik (for all the "five minutes" he dedicated to me), Liselotte, Leena, Lotta, Johan T., Ola, Christofer, Kristoffer, Aakash and, last but not least, Johan M., who first convinced me that there could be something useful in my 400 orbi runs.

Grazie alla comunità degli italiani in fuga, che mi ha fatto sentire più vicino a casa e mi ha insegnato molto: Cinzia, Roberto, Giorgio e Michele. Un ringraziamento inoltre alla Prof.ssa Sorgato, senza la quale non sarei arrivato a Lund.

Thanks also to all the people at Immunteknologi that I'm not explicitly mentioning here. I rarely fit coffee and fika in my day schedule but I did enjoy the breaks I shared with many people at the department. Thanks to Carl (even) for the Easter eggs.

# REFERENCES


Addona, T.A. et al., 2009. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nature Biotechnology*, 27(7), pp.633–641.

Aebersold, R. & Mann, M., 2003. Mass spectrometry-based proteomics. *Nature*, 422(6928), pp.198–207.

Allet, N. et al., 2004. In vitro and in silico processes to identify differentially expressed proteins. *PROTEOMICS*, 4(8), pp.2333–2351.

Anderson, N.G. & Anderson, N.L., 1982. The human protein index. *JAMA*, 246, pp.2620–2621.

Anderson, N.G., Matheson, A. & Anderson, N.L., 2001. Back to the future: the human protein index (HPI) and the agenda for post-proteomic biology. *PROTEOMICS*, 1(1), pp.3–12.

Anderson, N.L., 2010. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clinical chemistry*, 56(2), pp.177–185.

Anderson, N.L. & Anderson, N.G., 2002. The human plasma proteome: history, character, and diagnostic prospects. *Molecular & cellular proteomics : MCP*, 1(11), pp.845–867.

Anderson, N.L. et al., 2009. A human proteome detection and quantitation project. *Molecular & Cellular Proteomics*, 8(5), pp.883–886.

Anderson, N.L. et al., 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), pp.235–244.

Andersson, L. & Porath, J., 1986. Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. *Analytical biochemistry*, 154(1), pp.250–254.

Antoniou, A.C. & Easton, D.F., 2006. Models of genetic susceptibility to breast cancer. *Oncogene*, 25(43), pp.5898–5905.

Aryal, U.K. & Ross, A.R.S., 2010. Enrichment and analysis of phosphopeptides under different experimental conditions using titanium dioxide affinity chromatography and mass spectrometry. *Rapid Communications in Mass Spectrometry*, 24(2), pp.219–231.

Axelson, H. et al., 2005. Hypoxia-induced dedifferentiation of tumor cells--a mechanism behind heterogeneity and aggressiveness of solid tumors. *Seminars in cell & developmental biology*, 16(4-5), pp.554–563.

Beausoleil, S.A. et al., 2004. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(33), pp.12130–12135.

Boisvert, F.-M. et al., 2010. A quantitative proteomics analysis of subcellular proteome localization and changes induced by DNA damage. *Molecular & Cellular Proteomics*, 9(3), pp.457–470.

Bouwman, P. & Jonkers, J., 2012. The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance. *Nature Reviews Cancer*, 12(9), pp.587–598.

Brandsma, I. & Gent, D.C., 2012. Pathway choice in DNA double strand break repair: observations of a balancing act. *Genome integrity*, 3(1), p.9.

Brenton, J.D. et al., 2005. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(29), pp.7350–7360.

Brown, J.M. & Giaccia, A.J., 1998. The unique physiology of solid tumors: opportunities (and problems) for cancer therapy. *Cancer Research*, 58(7), pp.1408–1416.

Brunet, S. et al., 2003. Organelle proteomics: looking at less to see more. *Trends in cell biology*, 13(12), pp.629–638.

Castedo, M. et al., 2004. Cell death by mitotic catastrophe: a molecular definition. *Oncogene*, 23(16), pp.2825–2837.

Chapman, J.R., Taylor, M.R.G. & Boulton, S.J., 2012. Playing the end game: DNA double-strand break repair pathway choice. *Molecular Cell*, 47(4), pp.497–510.

Chi, A. et al., 2007. Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), pp.2193–2198.

Ciccia, A. & Elledge, S.J., 2010. The DNA damage response: making it safe to play with knives. *Molecular Cell*, 40(2), pp.179–204.

Claassen, M., 2012. Inference and validation of protein identifications. *Molecular & Cellular Proteomics*, 11(11), pp.1097–1104.

Cohen-Jonathan, E., Bernhard, E.J. & McKenna, W.G., 1999. How does radiation kill cells? *Current Opinion in Chemical Biology*, 3(1), pp.77–83. Available at: http://www.sciencedirect.com/science/article/pii/S1367593199800143.

Corthals, G.L. et al., 2000. The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis*, 21(6), pp.1104–1115.

Cox, Jürgen & Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), pp.1367–1372.

Cox, Jürgen et al., 2009. A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nature Protocols*, 4(5), pp.698–705.

Cox, Jürgen et al., 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research*, 10(4), pp.1794–1805.

Cox, Jürgen, Hubner, N.C. & Mann, M., 2008. How much peptide sequence information is contained in ion trap tandem mass spectra? *JAM*, 19(12), pp.1813–1820.

Craig, R. et al., 2006. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research*, 5(8), pp.1843–1849.

Craig, R., Cortens, J.P. & Beavis, R.C., 2004. Open source system for analyzing, validating, and storing protein identification data. *Journal of proteome research*, 3(6), pp.1234–1242.

Creasy, D.M. & Cottrell, J.S., 2004. Unimod: Protein modifications for mass spectrometry. *PROTEOMICS*, 4(6), pp.1534–1536.

De Bont, R., 2004. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis*, 19(3), pp.169–185.

DeSantis, C. et al., 2011. Breast cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(6), pp.409–418.

Deutsch, E.W., Lam, H. & Aebersold, R., 2008. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO reports*, 9(5), pp.429–434.

Domon, B. & Aebersold, R., 2006. Mass spectrometry and protein analysis. *Science*, 312(5771), pp.212–217.

Downs-Holmes, C. & Silverman, P., 2011. Breast cancer: overview & updates. *The Nurse practitioner*, 36(12), pp.20–6– quiz 7.

Duffy, M.J., Evoy, D. & McDermott, E.W., 2010. CA 15-3: uses and limitation as a biomarker for breast cancer. *Clinica chimica acta; international journal of clinical chemistry*, 411(23-24), pp.1869–1874.

Dumitrescu, R.G. & Cotarla, I., 2005. Understanding breast cancer risk -- where do we stand in 2005? *Journal of cellular and molecular medicine*, 9(1), pp.208–221.

Elston, C.W. & Ellis, I.O., 1991. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5), pp.403–410.

Eng, J.K. et al., 2011. A face in the crowd: recognizing peptides through database search. *Molecular & Cellular Proteomics*, 10(11), p.R111.009522.

Engelsberger, W.R. et al., 2006. Metabolic labeling of plant cell cultures with K(15)NO3 as a tool for quantitative analysis of proteins and metabolites. *Plant methods*, 2, p.14.

Fenn, J.B. et al., 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926), pp.64–71.

Fisher, B. et al., 2002. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *The New England journal of medicine*, 347(16), pp.1233–1241.

Galluzzi, L. et al., 2012. Molecular definitions of cell death subroutines: recommendations of the Nomenclature Committee on Cell Death 2012. *Cell death and differentiation*, 19(1), pp.107–120.

Geiger, T. et al., 2013. Initial quantitative proteomic map of twenty-eight mouse tissues using the SILAC mouse. *Molecular & Cellular Proteomics*.

Geiger, T., Cox, Juergen & Mann, M., 2010a. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Molecular & Cellular Proteomics*, 9(10), pp.2252–2261.

Geiger, T., Cox, Juergen, Ostasiewicz, P., et al., 2010b. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature Methods*, 7(5), pp.383–385.

Gerber, S.A. et al., 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12), pp.6940–6945.

Gillet, L.C. et al., 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*, 11(6), p.O111.016717.

Gordan, J.D. & Simon, M.C., 2007. Hypoxia-inducible factors: central regulators of the tumor phenotype. *Current opinion in genetics & development*, 17(1), pp.71–77.

Görg, A., Weiss, W. & Dunn, M.J., 2004. Current two-dimensional electrophoresis technology for proteomics. *PROTEOMICS*, 4(12), pp.3665–3685.

Gygi, S.P. et al., 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17(10), pp.994–999.

Haab, B.B., Dunham, M.J. & Brown, P.O., 2001. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome biology*, 2(2), p.RESEARCH0004.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–674.

Hanahan, D. & Weinberg, R.A., 2000. The hallmarks of cancer. *Cell*, 100(1), pp.57–70.

Hardman, M. & Makarov, A.A., 2003. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Analytical Chemistry*, 75(7), pp.1699–1705.

Havlis, J. & Shevchenko, A., 2004. Absolute quantification of proteins in solutions and in polyacrylamide gels by mass spectrometry. *Analytical Chemistry*, 76(11), pp.3029–3036.

Helczynska, K. et al., 2003. Hypoxia promotes a dedifferentiated phenotype in ductal breast carcinoma in situ. *Cancer Research*, 63(7), pp.1441–1444.

Hilger, M. et al., 2009. Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Molecular & Cellular Proteomics*, 8(8), pp.1908–1920.

Hilz, H., Wiegers, U. & Adamietz, P., 1975. Stimulation of proteinase K action by denaturing agents: application to the isolation of nucleic acids and the degradation of "masked" proteins. *European journal of biochemistry / FEBS*, 56(1), pp.103–108.

Hoopmann, M.R. & Moritz, R.L., 2013. Current algorithmic solutions for peptide-based proteomics data generation and identification. *Current opinion in biotechnology*, 24(1), pp.31–38.

Hörth, P. et al., 2006. Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. *Molecular & cellular proteomics : MCP*, 5(10), pp.1968–1974.

Hu, Q. et al., 2005. The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry : JMS*, 40(4), pp.430–443.

Hu, Zhi et al., 2010. The expression level of HJURP has an independent prognostic impact and predicts the sensitivity to radiotherapy in breast cancer. *Breast Cancer Research*, 12(2), p.R18.

Hu, Zhiyuan et al., 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics*, 7, p.96.

Huang, T. et al., 2012. Protein inference: a review. *Briefings in bioinformatics*, 13(5), pp.586–614.

Hunt, D.F. et al., 1986. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 83(17), pp.6233–6237.

Hüttenhain, R. et al., 2012. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. *Science translational medicine*, 4(142), p.142ra94.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945.

Ishihama, Y. et al., 2005. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & cellular proteomics : MCP*, 4(9), pp.1265–1272.

Jackson, S.P. & Bartek, J., 2009. The DNA-damage response in human biology and disease. *Nature*, 461(7267), pp.1071–1078.

Jansson, M. et al., 2008. Membrane protein identification: N-terminal labeling of nontryptic membrane protein peptides facilitates database searching. *Journal of proteome research*, 7(2), pp.659–665.

Jemal, A. et al., 2011. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2), pp.69–90.

Jiang, J. et al., 2007. Development of an immuno tandem mass spectrometry (iMALDI) assay for EGFR diagnosis. *Proteomics. Clinical applications*, 1(12), pp.1651–1659.

Jögi, A. et al., 2002. Hypoxia alters gene expression in human neuroblastoma cells toward an immature and neural crest-like phenotype. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), pp.7021–7026.

Kaiser, R. & Metzka, L., 1999. Enhancement of cyanogen bromide cleavage yields for methionyl-serine and methionyl-threonine peptide bonds. *Analytical biochemistry*, 266(1), pp.1–8.

Karas, M. & Hillenkamp, F., 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60(20), pp.2299–2301.

Kavanagh, J.N. et al., 2013. DNA Double Strand Break Repair: A Radiation Perspective. *Antioxidants & redox signaling*.

Kelleher, N.L. et al., 1999. Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry. *Journal of the American Chemical Society*, 121(4), pp.806–812.

Khatcheressian, J.L. et al., 2013. Breast cancer follow-up and management after primary treatment: american society of clinical oncology clinical practice guideline update. *Journal of Clinical Oncology*, 31(7), pp.961–965.

King, M.-C. et al., 2003. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, 302(5645), pp.643–646.

Krijgsveld, J. et al., 2003. Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nature Biotechnology*, 21(8), pp.927–931.

Lacey, J.M. et al., 2001. Rapid determination of transferrin isoforms by immunoaffinity liquid chromatography and electrospray mass spectrometry. *Clinical chemistry*, 47(3), pp.513–518.

Lacroix, M. & Leclercq, G., 2004. Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast cancer research and treatment*, 83(3), pp.249–289.

Lam, H., 2011. Building and searching tandem mass spectral libraries for peptide identification. *Molecular & Cellular Proteomics*, 10(12), p.R111.008565.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Lange, V. et al., 2008. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 4, p.222.

Langlands, F.E. et al., 2013. Breast cancer subtypes: response to radiotherapy and potential radiosensitisation. *The British journal of radiology*, 86(1023), p.20120601.

Lewanski, C.R. & Gullick, W.J., 2001. Radiotherapy and cellular signalling. *The lancet oncology*, 2(6), pp.366–370.

Link, A.J. et al., 1999. Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, 17(7), pp.676–682.

Liu, H., Sadygov, R.G. & Yates, J.R., 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, 76(14), pp.4193–4201.

Liu, L. et al., 2012. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 2012, p.251364.

Ludwig, J.A. & Weinstein, J.N., 2005. Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer*, 5(11), pp.845–856.

Ma, B. & Johnson, R., 2012. De novo sequencing and homology searching. *Molecular & Cellular Proteomics*, 11(2), p.O111.014902.

Ma, Y. et al., 2005. Repair of double-strand DNA breaks by the human nonhomologous DNA end joining pathway: the iterative processing model. *Cell cycle (Georgetown, Tex.)*, 4(9), pp.1193–1200.

Makinde, A.Y. et al., 2012. Radiation Survivors: Understanding and exploiting the phenotype following fractionated radiation therapy. *Molecular Cancer Research*.

Maris, J.M. et al., 2007. Neuroblastoma. *Lancet*, 369(9579), pp.2106–2120.

Marouga, R., David, S. & Hawkins, E., 2005. The development of the DIGE system: 2D fluorescence difference gel analysis technology. *Analytical and Bioanalytical Chemistry*, 382(3), pp.669–678.

Master, J.R.W., 2000. Human cancer cell lines: fact and fantasy. *Nature Reviews*, 1(DECEMBER 2000), pp.233–236.

McCabe, N. et al., 2006. Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. *Cancer Research*, 66(16), pp.8109–8115.

Mead, J.A., Shadforth, I.P. & Bessant, C., 2007. Public proteomic MS repositories and pipelines: available tools and biological applications. *PROTEOMICS*, 7(16), pp.2769–2786.

Miki, Y. et al., 2012. The advantages of co-culture over mono cell culture in simulating in vivo environment. *The Journal of steroid biochemistry and molecular biology*, 131(3-5), pp.68–75.

Milanowska, K. et al., 2011. REPAIRtoire--a database of DNA repair pathways. *Nucleic acids research*, 39(Database issue), pp.D788–92.

Mirgorodskaya, O.A. et al., 2000. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Communications in Mass Spectrometry*, 14(14), pp.1226–1232.

Misteli, T. & Soutoglou, E., 2009. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nature Reviews Molecular Cell Biology*, 10(4), pp.243–254.

Moeller, B.J. & Dewhirst, M.W., 2004. Raising the bar: how HIF-1 helps determine tumor radiosensitivity. *Cell cycle (Georgetown, Tex.)*, 3(9), pp.1107–1110.

Neve, R.M. et al., 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6), pp.515–527.

Neville, D.C. et al., 1997. Evidence for phosphorylation of serine 753 in CFTR using a novel metal-ion affinity resin and matrix-assisted laser desorption mass spectrometry. *Protein science : a publication of the Protein Society*, 6(11), pp.2436–2445.

Nielsen, T.O. et al., 2004. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10(16), pp.5367–5374.

O'Farrell, P.H., 1975. High resolution two-dimensional electrophoresis of proteins. *The Journal of biological chemistry*, 250(10), pp.4007–4021.

Oberle, C. & Blattner, C., 2010. Regulation of the DNA Damage Response to DSBs by Post-Translational Modifications. *Current genomics*, 11(3), pp.184–198.

Oda, Y. et al., 1999. Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12), pp.6591–6596.

Olsen, J.V. et al., 2006. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3), pp.635–648.

Olsen, J.V. et al., 2007. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9), pp.709–712.

Olsen, J.V., Ong, S.-E. & Mann, M., 2004. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & cellular proteomics : MCP*, 3(6), pp.608–614.

Olsson, N. et al., 2011. Proteomic analysis and discovery using affinity proteomics and mass spectrometry. *Molecular & Cellular Proteomics*, 10(10), p.M110.003962.

Ong, S.-E. & Mann, M., 2007. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature Protocols*, 1(6), pp.2650–2660.

Ong, S.-E. et al., 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), pp.376–386.

Ong, S.-E., Mittler, G. & Mann, M., 2004. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nature Methods*, 1(2), pp.119–126.

Paik, S. et al., 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine*, 351(27), pp.2817–2826.

Panchaud, A. et al., 2009. Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean. *Analytical Chemistry*, 81(15), pp.6481–6488.

Perkins, D.N. et al., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18), pp.3551–3567.

Perou, C.M. et al., 2000. Molecular portraits of human breast tumours. *Nature*, 406(6797), pp.747–752.

Picotti, P. et al., 2008. A database of mass spectrometric assays for the yeast proteome. *Nature Methods*, 5(11), pp.913–914.

Pinkse, M.W.H. et al., 2004. Selective isolation at the femtomole level of phospho-peptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Analytical Chemistry*, 76(14), pp.3935–3943.

Polanski, M. & Anderson, N.L., 2007. A list of candidate cancer biomarkers for targeted proteomics. *Biomarker insights*, 1, pp.1–48.

Prakash, A. et al., 2009. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. *Journal of proteome research*, 8(6), pp.2733–2739.

Puissant, A. et al., 2013. Targeting MYCN in Neuroblastoma by BET Bromodomain Inhibition. *Cancer discovery*.

Rahbar, A.M. & Fenselau, C., 2004. Integration of Jacobson's pellicle method into proteomic strategies for plasma membrane proteins. *Journal of proteome research*, 3(6), pp.1267–1277.

Razavi, M. et al., 2012. High-throughput SISCAPA quantitation of peptides from human plasma digests by ultrafast, liquid chromatography-free mass spectrometry. *Journal of proteome research*, 11(12), pp.5642–5649.

Roberts, G.C. & Smith, C.W.J., 2002. Alternative splicing: combinatorial output from the genome. *Current Opinion in Chemical Biology*, 6(3), pp.375–383.

Rodríguez-Ortega, M.J. et al., 2006. Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. *Nature Biotechnology*, 24(2), pp.191–197.

Ross, P.L. et al., 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP*, 3(12), pp.1154–1169.

Schena, M. et al., 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(20), pp.10614–10619.

Schindler, J. & Nothwang, H.G., 2006. Aqueous polymer two-phase systems: effective tools for plasma membrane proteomics. *PROTEOMICS*, 6(20), pp.5409–5417.

Scigelova, M. & Makarov, A., 2006. Orbitrap mass analyzer--overview and applications in proteomics. *PROTEOMICS*, 6 Suppl 2, pp.16–21.

Scrivener, E. et al., 2003. Peptidomics: A new approach to affinity protein microarrays. *PROTEOMICS*, 3(2), pp.122–128.

Semenza, G.L., 2003. Targeting HIF-1 for cancer therapy. *Nature Reviews Cancer*, 3(10), pp.721–732.

Shrivastav, M., De Haro, L.P. & Nickoloff, J.A., 2008. Regulation of DNA double-strand break repair pathway choice. *Cell research*, 18(1), pp.134–147.

Simpson, P.T. et al., 2005. Molecular evolution of breast cancer. *The Journal of Pathology*, 205(2), pp.248–254.

Simsek, D. et al., 2011. DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation. *PLoS genetics*, 7(6), p.e1002080.

Sorlie, T. 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), pp.8418–8423.

Sotiriou, C. et al., 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18), pp.10393–10398.

Speers, A.E. & Wu, C.C., 2007. Proteomics of integral membrane proteins--theory and application. *Chemical reviews*, 107(8), pp.3687–3714.

Steen, H. & Mann, M., 2004. The ABC“s (and XYZ”s) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5(9), pp.699–711.

Syka, J.E.P. et al., 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), pp.9528–9533.

Sørlie, T. 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), pp.10869–74.

Taguchi, A. & Hanash, S.M., 2013. Unleashing the power of proteomics to develop blood-based cancer markers. *Clinical chemistry*, 59(1), pp.119–126.

Tamatani, M. et al., 2000. A pathway of neuronal apoptosis induced by hypoxia/reoxygenation: roles of nuclear factor-kappaB and Bcl-2. *Journal of neurochemistry*, 75(2), pp.683–693.

Teschendorff, A.E. & Caldas, C., 2009. The breast cancer somatic "muta-ome": tackling the complexity. *Breast Cancer Research*, 11(2), p.301.

Thakur, S.S. et al., 2011. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular & Cellular Proteomics*, 10(8), p.M110.003699.

Thompson, A. et al., 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8), pp.1895–1904.

Ubersax, J.A. & Ferrell, J.E., 2007. Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology*, 8(7), pp.530–541.

Uehiro, N. et al., 2013. Validation study of the UICC TNM classification of malignant tumors, seventh edition, in breast cancer. *Breast Cancer ...*.

Unlü, M., Morgan, M.E. & Minden, J.S., 1997. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, 18(11), pp.2071–2077.

van 't Veer, L.J. et al., 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871), pp.530–536.

van de Rijn, M. et al., 2002. Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *The American journal of pathology*, 161(6), pp.1991–1996.

Van Hoof, D. et al., 2007. An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics. *Nature Methods*, 4(9), pp.677–678.

Vaupel, P. & Mayer, A., 2005. Hypoxia and anemia: effects on tumor biology and treatment resistance. *Transfusion clinique et biologique : journal de la Société française de transfusion sanguine*, 12(1), pp.5–10.

Venable, J.D. et al., 2004. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, 1(1), pp.39–45.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science*, 291(5507), pp.1304–1351.

Vizcaíno, J.A. et al., 2013. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research*, 41(Database issue), pp.D1063–9.

Vizcaíno, J.A., Foster, J.M. & Martens, L., 2010. Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. *Journal of proteomics*, 73(11), pp.2136–2146.

Wang, C. & Lees-Miller, S.P., 2013. Detection and Repair of Ionizing Radiation-Induced DNA Double Strand Breaks: New Developments in Nonhomologous End Joining. *International journal of radiation oncology, biology, physics*.

Ward, J.F., 1988. DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability. *Progress in nucleic acid research and molecular biology*, 35, pp.95–125.

Washburn, M.P., Wolters, D. & Yates, J.R., 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19(3), pp.242–247.

Weigelt, B., Geyer, F.C. & Reis-Filho, J.S., 2010. Histological types of breast cancer: how special are they? *Molecular oncology*, 4(3), pp.192–208.

Weinberg, R.A., 2007. The biology of cancer.

Weir, B., Zhao, X. & Meyerson, M., 2004. Somatic alterations in the human cancer genome. *Cancer Cell*, 6(5), pp.433–438.

White, P.S. et al., 1995. A region of consistent deletion in neuroblastoma maps within human chromosome 1p36.2-36.3. *Proceedings of the National Academy of Sciences of the United States of America*, 92(12), pp.5520–5524.

Whiteaker, J.R. et al., 2011. A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nature Biotechnology*, 29(7), pp.625–634.

Wilkins, M.R. et al., 1999. High-throughput mass spectrometric discovery of protein post-translational modifications. *Journal of molecular biology*, 289(3), pp.645–657.

Woolston, C.M. et al., 2011. Expression of thioredoxin system and related peroxiredoxin proteins is associated with clinical outcome in radiotherapy treated early stage breast cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, 100(2), pp.308–313.

Wu, C.C. et al., 2004. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Analytical Chemistry*, 76(17), pp.4951–4959.

Yarnold, J., 2009. Early and locally advanced breast cancer: diagnosis and treatment National Institute for Health and Clinical Excellence guideline 2009. *Clinical oncology (Royal College of Radiologists (Great Britain))*, 21(3), pp.159–160.

Zhang, Z. & Chan, D.W., 2010. The road from discovery to clinical diagnostics: lessons learned from the first FDA-cleared in vitro diagnostic multivariate index assay of proteomic biomarkers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 19(12), pp.2995–2999.

Zhao, Y. & Jensen, O.N., 2009. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *PROTEOMICS*, 9(20), pp.4632–4641.

*Raffaello, "Scuola di Atene", detail* (from Wikimedia Commons).