Translational Bioinformatics 6
Series Editor: Xiangdong Wang, MD, PhD, Prof

György Marko-Varga *Editor*

Genomics and Proteomics for Clinical Discovery and Development

Springer

Translational Bioinformatics

Series editor

Xiangdong Wang, MD, Ph.D.
Professor of Clinical Bioinformatics, Lund University, Sweden
Professor of Medicine, Fudan University, China

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Recently Published and Forthcoming Volumes

Applied Computational Genomics
Editor: Yin Yao Shugart
Volume 1

Pediatric Biomedical Informatics
Editor: John Hutton
Volume 2

Bioinformatics of Human Proteomics
Editor: Xiangdong Wang
Volume 3

Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases
Editor: Bairong Shen
Volume 4

More information about this series at
http://www.springer.com/series/11057

György Marko-Varga
Editor

Genomics and Proteomics for Clinical Discovery and Development

Springer

Editor
György Marko-Varga
Clinical Protein Science and Imaging
Biomedical Center
Biomedical Engineering
Lund University
Lund, Sweden

Center of Excellence in Biological
and Medical Mass Spectrometry
Biomedical Center D13
Lund University
Lund, Sweden

First Department of Surgery
Tokyo Medical University
Tokyo, Japan

ISSN 2213-2775 ISSN 2213-2783 (electronic)
ISBN 978-94-017-9201-1 ISBN 978-94-017-9202-8 (eBook)
DOI 10.1007/978-94-017-9202-8
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2014946410

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The developments within the protein and gene-expression area, where proteomics, also commonly addressed as the post-genomic research field is growing rapidly, with ever increasing value and power within life sciences. The current book on *Genomics and Proteomics for Clinical Discovery and Development* will aid in the understanding and utilization of these new research areas that are expanding rapidly.

There is a natural interaction today between academic, industrial, biotech, and pharmaceutical colleagues who share the common interest in developing our understanding of the protein components within pathways, systems, and organisms. We are a community interested in providing detailed knowledge about the structure-function-expression inter-relationships that support and determine cellular activity. We have a strong link between technology and biology which gives us an unprecedented opportunity to understand the natural history of disease processes. Our levels of understanding and our appreciation of the complexities surrounding protein expression, function, and detection have grown with sophistication in applications of technology and a genuine understanding of the interactive networks of protein modules and pathways present at every level of biology.

The impact of the field in terms of deliveries within the drug discovery and the drug development area globally is significant. In addition, the diagnostic field is progressing in an impressive speed. So do the new developments of instrument and technology platforms that currently have a billion dollar turnaround.

The various parts of the book have been divided into parts that will cover the most expanding and most important areas within proteomics. The result is this volume.

We thank all of the authors and contributors and the Springer publisher for support throughout the production phase and final editorial processing, and for sponsoring this dedicated issue. We hope that it will serve as a platform for discussion and debate. We intended that the individual papers could serve as a primer for students and Professors, for Physicists and Deans of Medicine, and for all of us within the proteomics community. Finally, we wish to acknowledge and express our sincere gratitude to the authors and contributors of the chapter that follows.

The different chapters are presented by well-recognized scientists and specialists within the respective field of genomics and proteomics. The authors they come from vary by different countries and continents, and are representatives from both the healthcare sector, academia, pharmaceutical industry and biotech industry. In other words this constitutes a well-mixed expert team.

The different parts of this book include a survey of all human proteins coded by the human genome, and the assignments of unknown proteins. The transcript profiling, whereby the unknown proteins can be identified and annotated. The standardization of protein annotations that follow a guideline, declared and agreed by the international science community, is also presented and examples are given by experts in the field. Mass spectrometry is the key technology for sequencing and analyzing proteins, and this part is outlined in detail, presenting basic analysis procedures, as well as algorithms and procedures for data analysis on a number of mass spectrometry technology platforms. These platforms are then applied to the discovery of biomarkers in cardiology as well as oncology, where post-translational modifications are outlined in the respective studies.

High density protein microarrays utilizing NAPPA technology is a powerful principle of screening large number of proteins and samples. The bioinformatics elucidation of data generated by proteomics linking it to bio-statistical analysis is also an important part that builds the proteomic studies. Here the link to epigenetics is outlined and extensive examples are illustrated in a dedicated chapter.

Drug development is a major part of proteomics and genomics applications, and the pharmaco-genetics link in these developments is outlined. In order to provide the best pipeline for development, patient samples of high quality generated by Biobanks is a resource that currently is expanding heavily around the world.

The drug localization and mode of drug action is important in order to be able to provide evidence of the mechanism of the drug whereby it interferes with the pathology. Imaging techniques are currently under fast development and their utilities are outlined in a dedicated chapter of the book.

Finally, I would like to express my sincere appreciation to all of the authors, for their invaluable scientific contributions that made this book project possible.

Lund, Sweden György Marko-Varga

Contents

Chapter 1
Introduction to Genomics and Proteomics for Clinical Discovery and Development

György Marko-Varga

Abstract In the post-genomic era, Genomics and Proteomics has become the major OMICS research areas utilised to understand biological processes and disease pathophysiology that leads to drug development and diagnostic assays.

Proteome-based approaches are important complements to genomic data and provide crucial information of the target driver molecules and their post-translational modifications, where informatics and standardizations are of key importance.

The Chromosome initiative of the Human Proteome Project currently annotates the number of proteins that are coded by the genome. This chapter gives an introduction to Genomics and Proteomics for Clinical Discovery and Development.

Keywords Cardiovascular disease • MRM • Protein sequencing • Proteomics • Genes • mRNA • Mass spectrometry

1.1 Genomics and Proteomics for Clinical Discovery and Development

There is a highly unmet need within the healthcare sector that calls for an increasing worldwide demand for new medicines and treatments. This in order to utilize and apply cutting edge of research with state-of-art facilities and opportunity that can be used in order to impact on quality of life of people. By entering into the post-genomic era, it has become evident that genetic changes alone are not sufficient to understand most biological and pathophysiological processes. Throughout the last decades,

G. Marko-Varga (✉)
Clinical Protein Science and Imaging, Biomedical Center,
Biomedical Engineering, Lund University, BMC D13, 221 84 Lund, Sweden

Center of Excellence in Biological and Medical Mass Spectrometry,
Biomedical Center D13, Lund University, 221 84 Lund, Sweden

First Department of Surgery, Tokyo Medical University,
6-7-1 Nishishinjiku Shinjuku-ku, Tokyo 160-0023, Japan
e-mail: gyorgy.marko-varga@bme.lth.se

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_1

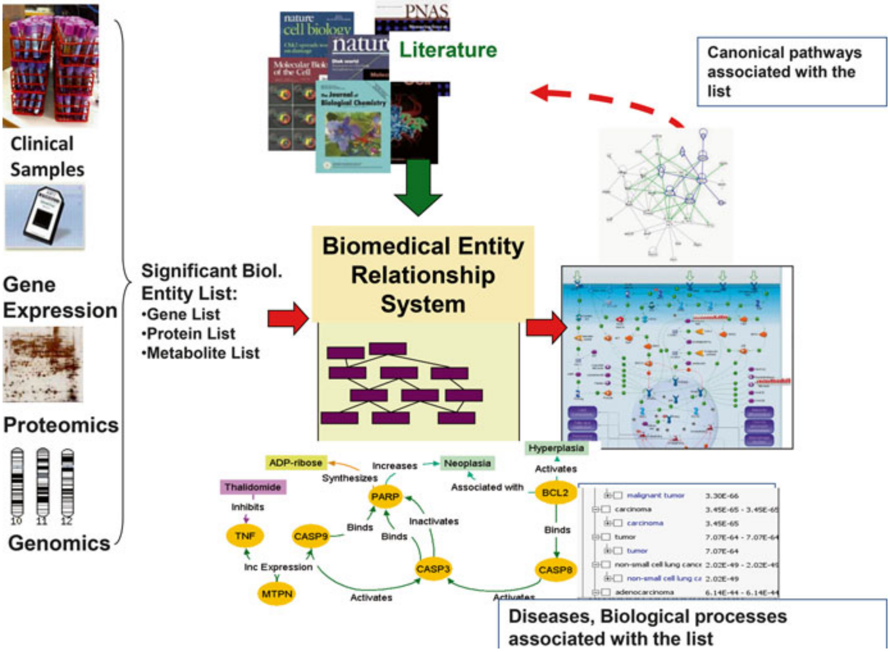


Fig. 1.1 The drug and diagnostic work flow; Schematic overview of the drug and diagnostics developments work flow where the protein-, and gene sequencing platforms are key in the process that also involve the utilization of available literature and text mining. Interfacing in-between the process steps is mandatory in order to generate candidates and the finally result is clinical products that can be brought to the market and used by patients

Genomics has publicized a large number of Genomes, included in the work of the 1000 genomes project (http://www.1000genomes.org/). Proteomics and genomics are the two main areas within the OMICS research field that has grown with an ever increasing speed over the last two decades. These areas can be defined in many different ways, depending on the range of disciplines that focuses on the "Life Science" target of protein-, and gene sequencing that forms the basis for drug and diagnostics developments, as outlined in Fig. 1.1.

The Proteomics start can be traced back to mid-1990 when the term "Proteome" and Proteomics was first mentioned and introduced by Wilkins and Williams. Two decades later, Proteomics finds itself as a highly integrated field of research with a highly versatile and efficient way of pursuing medical and biological science at an unprecedented pace.

In comparison to Genomics, Proteomics is a much more complex area, as the building blocks of the amino acids are fife-fold higher. Another difference that has impacted the development of Proteomics is the lack of a PCR technique available. The lack of amplification upon detection, is a major challenge in many protein expression studies e.g., in disease areas, where the sensitivity gain, only can be accomplished by using increasing amounts of the sample material in the experiments. Another feature that adds to the Proteomics complexity is the post-translational modifications (PTMs) that occur in protein processing mechanisms. The PTM occurrence in a cellular context will also change the 3D-structure and composition of the protein, adding another dimension of complexity to the analysis within Proteomics. The technology available for studying proteome expression and resolving exact protein and peptide identities in complex mixtures of biological samples allows global protein expression within cells, fluids, and tissue to be approached with confidence.

The proteomics research areas within biomedical applications roughly segregate into four fields of categories of protein study:

- Expression proteomics
- Functional proteomics
- Structural proteomics
- Chemi-proteomics.

Mass Spectrometry has been the protein sequencing work horse engine for the last 15 years and has made considerable improvements over time, with increase in both resolution and sensitivity.

Today, it is possible for instance to sequence the yeast genome (roughly 2,000 proteins), within 1 h using the latest MS-instrument platform. These research areas are mainly driven by challenges in everyday medicine and the shortcoming of today's healthcare in close linkage to technology development. Consequently, the developments within both technology and medicine forms a yin-yang relationship, where the medical field is not able to manage without the revolutionary achievements in technology and vice versa.

1.2 The ENCODE Project

The ENCODE initiative is a NIH driven and funded Genomic program, that was launched in 2003. ENCODE is the successor of the Human Genome Project (HUGO).

After the announcement of the HUGO outcome in 2000, mapping all human genes, the ENCODE's goal is to identify all the functional elements of the human genome, This will include all of the 21,000 genes that make up a mere 1 % of its three billion nucleotides.

In 2012, the ENCODE consortium did announce the first comprehensive results of a 6-year-long effort publishing altogether 33 research papers in five journals, including *Nature* and *Science*.

In these milestone studies, ENCODE's signal claim, study data enables to assign biochemical functions for approx., 76 % of the genome. This includes outcomes that includes regions outside of the well-studied protein-coding regions (Lane et al. 2014; ENCODE Project Consortium 2012).

1.3 HPP – The Human Proteome Project

The HPP initiative is a Chromosome-centric Human Proteome Project: A New Systematic Approach to Catalogue Proteins Using the Integrated Genomics/ Proteomics Technologies (Paik et al. 2012a, b).

The Human Chromosome Initiative global research activities is developing for tools that can bridge the data sets of information held with each clinical sample with substantive deliverables that can be used to create future standards of health care (Marko-Varga et al. 2013; Paik and Hancock 2012).

With a well characterized human genome map together with the availability of in depth transcriptomics, the C-HPP came from the understanding that the proteomic community was well placed to study the full complexity of human proteome.

The objectives and goals of such a large-scale initiative is that it will expand our knowledge of the phenotypic state within biological and clinical research.

The C-HPP consortia teams have as a primary goal that the proteomic catalog should be put in the context of the chromosomal gene sequences to promote more effective collaborations with molecular biologists and to improve understanding of the biological context of proteomics data sets (Lane et al. 2014; ENCODE Project Consortium 2012; Paik et al. 2012a). As the human genome project (HGP) produced whole genomic parts list, the primary goal of C-HPP is to define the protein parts list of each chromosome by the effort of the 24-constituted national and international teams (Paik and Hancock 2012).

1.4 Personalized Medicine

In order to utilize a targeted drug "Personalized Medicine" an optimization is needed in order to select the correct patient group (Mok et al. 2011). Appropriate biomarkers are being used to select treatments best suited for the individual patient. The disease presentation will align the patient with the correct phenotype.

Recent discoveries that involves the drug developments towards somatic mutations and of activating mutations in *EGFR* and fusion genes was presented upon with mechanistic elucidations (Kirk 2012).

This study involved the ALK target that has set the stage for personalized medicine within the lung cancer patient cohorts, especially in Japan. Novel biomarkers have utilized by the patients, that benefited from the development of EGFR tyrosine kinase inhibitors (Gefitinin; "IRESSA" and Erlotinib "TARCEVA") and ALK inhibitors with a significant outcome that showed a clear improvement in both the tumor control as well as survival.

The key features of state-of-the-art proteomic expression profiling is to provide high resolution capacity with sensitivity that allows low abundant proteins to be assigned (Marko-Varga et al. 2007). This is of mandatory importance as the cancer therapy is moving toward individually selected treatments.

This emerging targeted therapy approach addresses both genotype and phenotype information, that includes protein expression. The effective therapy of Personalized Medicine is directly linked to the development of effective biomarker assay kits that provide predictive read-out of the response to treatment.

1.5 Biobanking

Large sample collections that represents patients disease status is of central importance in OMICS research. This was recognised recently by the TIME Magazine (Welinder et al. 2013), highlighted biobanking as One of "10 Ideas Changing the World Right Now."!

The Biobanking development is rapidly growing at a global scale and thriving to change approaches to target-finding, drug development and patient treatment. Biobanking is being viewed as a Key Driver for Next Generation Biomarker and Drug Discovery, including developments within Biobank initiatives that are quickly increasing Globally (Marko-Varga 2013).

Biobanks are archives of clinical sample materials from the Health Care area, where the use and outcome will aid in developments (Marko-Varga et al. 2012).

Applying the samples to Proteomics analysis aims at correlating the specific protein expression patterns of individual subjects with those of other patients. It may be possible to assign a statistical relationship of significance to specific protein expression patterns with other defined measurements of clinical presentation (Malm et al. 2012).

References

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

Kirk R. Genetics: personalized medicine and tumour heterogeneity. Nat Rev Clin Oncol. 2012;9:250.

Lane L, Bairoch A, Beavis RC, Deutsch EW, Gaudet P, Lundberg E, Omenn GS. Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. J Proteome Res. 2014;13(1):15–20.

Malm J, Végvári A, Rezei M, Upton P, Danmyr P, Nilsson R, Steinfelder E, Marko-Varga G. Large scale biobanking of blood – the importance of high density processing procedures. J Proteomics. 2012;76:116–24.

Marko-Varga G. BioBanking as the central tool for translational medicine CTM issue 2013. Clin Trans Med. 2013;2:4. doi:10.1186/2001-1326-2-4.

Marko-Varga G, Ogiwara A, Nishimura T, Kawamura T, Fujii K, Kawakami T, Kyono Y, Tu HK, Anyoji H, Kanazawa M, Akimoto S, Hirano T, Tsuboi M, Nishio K, Hada S, Jiang H, Fukuoka M, Nakata K, Nishiwaki Y, Kunito H, Peers IS, Harbron CG, South MC, Higenbottam T, Nyberg F, Kudoh S, Kato H. Personalized medicine and proteomics: lessons from non-small cell lung cancer. J Proteome Res. 2007;6:2925–35.

Marko-Varga G, Végvári A, Welinder C, Rezei M, Edula G, Svensson K, Belting M, Laurell T, Fehniger TE. Standardization and utilization of biobank resources in clinical protein science with examples of emerging applications. J Proteome Res. 2012;11:5124–34.

Marko-Varga G, Omenn GS, Paik YK, Hancock WS. A first step toward completion of a genome-wide characterization of the human proteome. J Proteome Res. 2013;12(1):1–5.

Mok TSK, et al. Personalized medicine in lung cancer: what we need to know. Nat Rev Clin Oncol. 2011;8:661–8.

Paik YK, Hancock WS. Uniting ENCODE with genome-wide proteomics. Nat Biotechnol. 2012;30(11):1065–7.

Paik Y-K, Jeong S-K, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee H-J, Na K, Choi E-Y, Yan F, Zhang F, Zhang Y, Snyder M, Cheng Y, Chen R, Marko-Varga G, Deutsch EW, Kim H, Kwon J-Y, Aebersold R, Bairoch A, Taylor AD, Kim KY, Lee E-Y, Hochstrasser D, Legrain P, Hancock WS. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotech. 2012a;30:221–223. http://www.nature.com/nbt/journal/v30/n3/abs/nbt.2152.html#supplementary-information.

Paik Y-K, Omenn GS, Uhlen M, Hanash S, Marko-Varga G, Aebersold R, Bairoch A, Yamamoto T, Legrain P, Lee H-J, Na K, Jeong S-K, He F, Binz P-A, Nishimura T, Keown P, Baker MS, Yoo JS, Garin J, Archakov A, Bergeron J, Salekdeh GH, Hancock WS. Standard guidelines for the Chromosome-Centric Human Proteome Project. J Proteome Res. 2012b;11:2005–13. doi:10.1021/pr200824a.

Welinder C, Jönsson G, Ingvar C, Lundgren L, Olsson H, Breslin T, Végvári A, Laurell T, Rezeli M, Jansson B, Baldetorp B, Marko-Varga G. Establishing a Southern Swedish Malignant Melanoma OMICS and biobank clinical capability. Clin Trans Med. 2013;2(1):7. doi:10.1186/2001-1326-2-7.

Chapter 2
Identification of Missing Proteins: Toward the Completion of Human Proteome

Ákos Végvári

2.1 Introduction

Undoubtedly, one of the greatest achievements of scientific research was the completion of sequencing the human genome in 2003 that is still unmatched in size of collaborative efforts (International Human Genome Sequencing Consortium 2004). The social benefit of the knowledge generated is better understanding the expressions of genes in healthy and diseased conditions, which in turn can lead to better diagnosis and treatment of many diseases, including various forms of cancer. Notably, the technological developments have continuously delivered efficient tools to overcome the difficulties of mapping about three billion base pair long genetic codes. As an additional outcome, today, we know that the human chromosomes hold about 20,300 genes coding for all functional proteins in the wide versatility of biological activities of cells.

The Human Genome Project has provided a blueprint of international collaborations for novel research programs, such as the recent Human Proteome Project (HPP) (http://www.hupo.org/initiatives/human-proteome-project/) organized by the Human Proteome Organization (HUPO) that was launched in September 2010 in Sydney at the HUPO World Congress. HPP has the aim to map the entire human proteome utilizing current technologies with mass spectrometry (MS), antibodies and knowledgebase in focus with respect to protein abundance, distribution, subcellular localization, interaction with other biomolecules and functions at specific time points. However, the number of human proteins assumed to greatly exceed the number of human genes because of some eukaryote specific phenomena, such as alternative splicing, and substantial post-translational modifications of amino

Á. Végvári, Ph.D. (✉)
Clinical Protein Science & Imaging, Department of Biomedical Engineering,
Lund University, Biomedical Center D13,
SE-211 84 Lund, Sweden
e-mail: akos.vegvari@bme.lth.se

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_2

acid side chains with various chemical groups. Currently, HPP has initiated two, complementary international collaborations, that are divided into the Chromosome-centric HPP (C-HPP) dealing with proteins coded on particular chromosomes and the Biology/Disease-Driven HPP (B/D-HPP) concentrating on the functions of human proteins.

Key elements of success in a research project of this measure are the utilization of latest technological advances, such we fortuned to appreciate in case of mass spectrometry and antibody based proteome analyses. Additionally, the variety of data has to be shared in order to evaluate, understand and make it use when it comes to clinical applications. Large sets of data have already been published in repositories and submitted to databases like the widely acknowledged Swiss Institute of Bioinformatics projects, UniProtKB/SwissProt and lately the golden standard neXt-Prot (Lane et al. 2012).

2.2 Current State of Human Protein Project

The purpose of the international C-HPP is to map and identify all human proteins that are encoded by genes localized on chromosomes using selected human samples (Paik et al. 2012a; Legrain et al. 2011). In addition to organize a joint network of research groups, C-HPP is determined to capture particular biological features of gene variation, gene regulation and protein expression coordinated at the chromosomal level (Paik et al. 2012a; Legrain et al. 2011). Accordingly, C-HPP project is generating and reporting data in a format that is aligned with the DNA sequence of individual chromosomes and with the output of transcriptome data (RNA sequencing). In addition, proteins will be characterized for major proteoforms resulted in by alternative splicing transcript (AST), non-synonymous single-nucleotide polymorphism (nsSNP) and post-translational modifications (PTMs). In order to meet these goals, a novel network of bioinformatics tools is about to be rendered including existing platforms such as PRIDE (http://www.ebi.ac.uk/pride/), GPMDB (http://gpmdb.thegpm.org), PeptideAtlas (http://www.peptideatlas.org), UniProt/SwissProt (http://www.uniprot.org) and emerging initiatives, such as NextProt (http://www.nextprot.org), ProteomeXchange (http://www.proteomexchange.org), Tranche (https://proteomecommons.org/tranche/) and HUPO Proteomics Standards Initiative (http://www.psidev.info) (Paik et al. 2012b). In parallel, the sibling program B/D-HPP is set to provide complementary knowledge from studies of cellular mechanisms and biochemical processes analyzing proteomes from the point of view of human diseases (Aebersold et al. 2013). This can facilitate life science research supporting routine determinations of processes and disease relevant proteins.

In this context, the first milestone delivery of HPP is the completion of the human protein catalogue, *i.e.*, the complete list of consensus proteins sequences. The consensus or canonical sequences are those proteoforms that occur most frequently in nature, coded often by the wild type allele of the given genes or longest protein sequence available. Currently, protein sequential data is collected at high pace

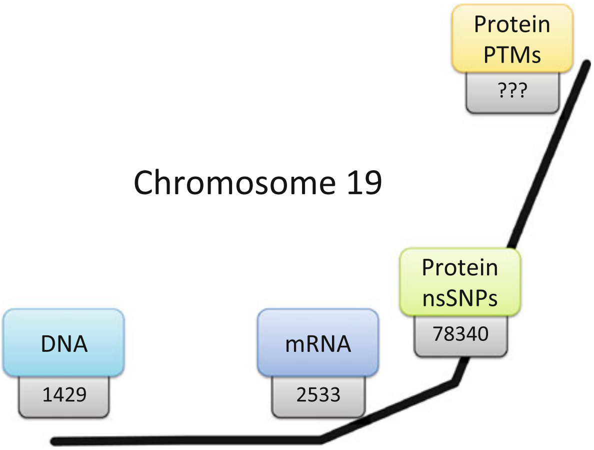


Fig. 2.1 Illustration of the number of proteins coded in chromosome 19 genes (DNA) reflecting the known and expected proteoforms following alterative splicing (mRNA), non-synonymous single nucleotide polymorphism (nsSNP) and post-translational modification (PTMs, without number)

mostly due to the technological advances in mass spectrometry. However, a substantial portion of the consensus sequences is yet unknown due to the lack of quality observations of a given protein, such as incorrect gene annotation, very low abundance, absence of expression in a certain tissue, expression only in rare samples, or unfavorable structure (or cleavage sites) for MS studies such as heterogeneity and instability.

Adding to the complexity of the task, proteins exist in multiple proteoforms including AST products, mutant variants (nsSNPs) and variable PTMs. As a consequence, the exact size of the human proteome is still not known today but may be expected to be several millions (Legrain et al. 2011; Lane et al. 2014). Since HPP is determined to identify at least one AST and one nsSNP product of each canonical protein sequence as well as three major PTMs (phosphorylation, glycosylation and acetylation) (Paik et al. 2012b), the number of entries in the complete human proteome planned is estimated in the range of 100,000–1,000,000. Figure 2.1 illustrates the multiple factors that boost the expected protein sequences keeping the focus on chromosome 19 only that represent about 14 % of the human genes coding proteins.

According to the common annotation procedure employed to curate protein databases (*e.g.*, UniProtKB/SwissProt), identification is defined at five levels: (1) evidence at protein level, (2) evidence at transcript level, (3) inferred from homology, (4) predicted and (5) uncertain. Importantly, protein level annotations require clear experimental evidence for the existence of the protein, including partial or complete Edman sequencing, unambiguous MS identification, X-ray or NMR structure, good quality protein-protein interaction or detection of the protein by antibodies. This category system was also introduced to neXtProt that is relied on the foundation of the large efforts made by SwissProt to functionally annotate human proteins and curate their sequences. Adapting to the goals of C-HPP, “missing proteins” were defined as those proteins without protein level identification that are (1) identified in

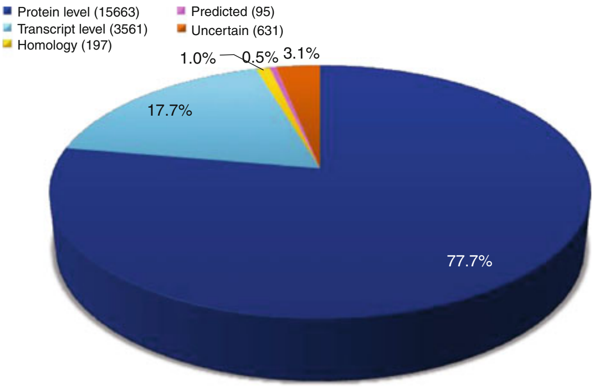


Fig. 2.2 The current state of the human proteome is shown as neXtProt data presented by release 2013-12-09 (http://www.nextprot.org)

transcriptome, (2) with a predicted or homology inferred sequence, (3) partially identified proteins with transcript evidence but without convincing MS information. As such, the primary goal of C-HPP is defined to compile the list of "missing proteins" (Paik et al. 2012a, b).

The reason for not being able to identify proteins is multifold, as Lane et al. has pointed out (Lane et al. 2014). Many proteins are expressed in unusual organs or cell types only leaving their observations a unique event. Notably, several genes common among many mammals are not expressed in humans as the olfactory receptor genes exemplify. Presumably, a significant number of proteins may only be expressed in early developmental stages in the embryo or fetus but relatively few proteins were reported so far (Munoz et al. 2011). An additional portion of genes and proteins may be silent and could be activated under certain stresses. Furthermore, high sequence homology in certain protein families may generate overlapping matches to available tryptic peptides making the identification of their origin impossible. Lastly, we should also mention the detection of limit of present analytical platforms as a potential source of "missing" proteins.

C-HPP has reported on the initial number of "missing proteins" about 30 % in 2011 (Legrain et al. 2011) that has melted down to about 22 % according to the latest release of the golden standard database of neXtProt (version 2013-12-09) as shown in Fig. 2.2. The proposed strategy to target proteins, which have been previously identified at transcript level, is supported by the fact that the largest number of "missing protein" falls into this category (17.7 %).

2.3 Methods for Identification

A variety of measurement approaches are currently in use for protein annotations. Fulfilling the SwissProt inclusion criteria, mentioned in the previous section, protein sequencing data required that is typically generated by MS nowadays. The C-HPP requested to map all proteins presently lacking high quality mass spectrometry confirmation, covering also three major classes of PTMs, characteristic AST products and nsSNP sequence variants. Further verifications should be followed by antibody-based detection in selected tissues and cell lines. To guide the selection of appropriate samples, as well as the preparation of recombinant protein standards and heavy-labeled proteotypic peptides in SRM assays transcript data generated by RNA sequencing is advantageous (Picotti et al. 2010).

Recently, a useful search strategy was proposed to identify "missing" proteins in human samples (Lane et al. 2014): (1) start from the tissue distribution of reported transcriptome expression; (2) consider early stages of life (3) consider special stresses or other perturbations; (4) recognize low abundance of many proteins or transmembrane helical structures and sequences poor in tryptic cleavage sites, requiring more sensitive or different analytical methods; (5) seek more detailed information about highly homologous families of proteins and increase the sequence coverage, if feasible (Lane et al. 2014).

During the last years, the protein science research field has developed new and complementary technologies, like protein shotgun sequencing and quantitative mass spectrometry platforms that are more and more widely available (Mann 2009; Olsen et al. 2009; Schmidt et al. 2009). Shotgun proteomics has generated high quality data facilitating identification of "missing proteins" with proteotypic tryptic peptides sequences in human samples. Since human biofluid is a rich source of patient sample for clinical analysis, healthcare has given protein quantitation in blood much attention over the last decades. However, plasma and serum display a concentration range of 10–12 orders of magnitude that is an enormous challenge for present analytical platforms today (Anderson and Anderson 2002; Domon and Aebersold 2006). Currently, most laboratories build protein assays functional only in 3–5 orders of magnitude of concentration and combine various measuring platforms to expand the range of interest.

One of the most popular mass spectrometry technology, selected reaction monitoring (SRM) can be successfully applied to identify and quantify specific peptides within the digested samples of complex mixture. In addition, the SRM methodology is inherently easy to multiplex allowing multiple protein analysis assays straightforward to develop that offer high sensitivity and speed. By the use of isotope labeling technology, uniformly ^{13}C-^{15}N-labeled blood plasma levels of 100 ng/mL proteins can be quantified.

However, in many cases additional enrichment steps can result in identification of proteins present at lower concentrations in human samples like plasma or serum. Targeted enrichment has been performed with or without antibodies to improve the detection sensitivity. Recently, several approaches combining immunoaffinity with SRM using stable isotope peptides, such as SISCAPA-SRM (Anderson et al. 2004), immuno-SRM (Whiteaker et al. 2007, 2011), mass spectrometric immunoassay (MSIA) (Lopez et al. 2010), have significantly improved the limit of detection of low abundant protein biomarkers present in plasma. Using the MSIA method, the lowest detection level (LOQ) of plasma proteins is 16–31 pg/mL (Lopez et al. 2010). Since antibodies are not always available and expensive to develop, antibody-free enrichment of target proteins was recently demonstrated for the quantitation of low abundant plasma proteins at concentrations in the 50–100 pg/mL range (Shi et al. 2012).

Additionally, RNA sequence analysis should be performed with biological samples and used the transcript data to verify newly identified proteins as well as known proteins for the given specimens. This approach can provide considerable reference information on each missing protein in a given sample. Upon verification, analyses of "missing" proteins can then be performed using recombinant proteins and mass spectrometry to produce peptide signatures that can be used for cross-checking with SRM library. Validation of protein identifications then can be further improved by antibodies at the cellular or molecular level using Western blotting or immune-cytological analyses employing antibodies raised against the synthetic peptides of the proteins of interest.

However, clearly this is not enough to populate a resource that needs to address the complexity of the universe of human proteins. Therefore, novel technologies and tailored approaches are expected to emerge in the field. An interesting procedure was recently proposed that presents a pipeline integrating a suite of bioinformatics analysis and annotation software tools targeting protein homologues and mapping putative functional signatures, gene ontology, and biochemical pathways (Islam et al. 2014). This approach could successfully annotate a set of "missing" proteins, finding that 33.2 % of them were homologous to nonhuman reviewed mammalian proteins with proteomic evidence.

2.4 Databases

One of the three basic principles, "pillars" (Legrain et al. 2011), of the proposed HPP workflow is knowledge base that may be regarded yet another methodology employing bioinformatic platforms and organizing experimental data in the form of databases. Systematic arrangements of proteomic data provides another level knowledge that comes to extremely useful for D/B-HPP studies. Because of the large amount of high quality proteomic data generated prior to the start of C-HPP, the first logical phase of the strategy was data integration and reciprocal validation with protein and translating evidence to explore the expression of coding genes (Zhong et al. 2014). Obviously, this approach primarily relies on knowledge base and tools to generate, access and integrate the accumulated information.

From this point of view databases serve as fundamental to search for missing proteins and to create a comprehensive list using several databases (*e.g.*, Ensembl, neXtProt, UniProt/SwissProt, GPMDB) by crosschecking the entire list of genes. Figure 2.3 shows the relationships between the core databases as outlined by HPP. C-HPP has selected those databases that are consistently play a central role during the entire project, including UniProtKB/SwissProt (The UniProt Consortium 2011), PRIDE (Martens et al. 2005), PeptideAtlas (Desiere et al. 2006), GPMDB (Fenyo et al. 2010) and Human Protein Atlas (Berglund et al. 2008) databases as

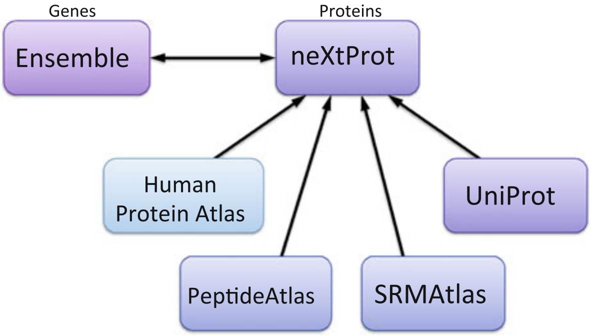


Fig. 2.3 Relationship between the core databases recommended by HPP

well as the ProteomeXchange (Craig et al. 2004) infrastructure for coordinating the proteomics databases through the Tranche (Smith et al. 2011) file-sharing system. Another knowledge platform, neXtProt (Lane et al. 2012; Gaudet et al. 2013), is pointed to be the golden standard that can connect the results of the HPP projects, in particular of MS and antibody based results with completed functional and structural database about all human proteins. The raw mass spectral data intended for protein identifications is requested to deposit with full annotation in Tranche, while datasets are shared with PeptideAtlas automatically. The SRM data generated with synthetic peptides are deposited in the SRMAtlas, whereas the knowledge about antibodies is gathered in the Human Protein Atlas and Antibodypedia (Björling and Uhlén 2008).

Unfortunately, complementary information about technical (inappropriate protein physical properties or tryptic digestion) and gene identification difficulties responsible for lacking annotation is not encompassed in the appropriate databases. Furthermore, protein databases, including neXtProt, provide the level of identification of the consensus sequence but do not indicate this in case of the AST and nsSNP products. This knowledge should be implemented in the key data resources since it is of great value for many researchers interested in knowing what proteoforms have been identified and characterized in various cell types, organs and biofluids (Lane et al. 2014). Since the most productive shotgun proteomic technology, delivering sequencing of peptides is relied on the consensus protein sequences in databases, it should be noted that “the human genome” is also a consensus sequence. As such, individual variations of many genes and their protein products might have been observed in high quality mass analyses but this evidence for certain proteoforms are largely neglected at this stage of HPP. It is greatly desired to extend both genome (Ensembl) and proteome (neXtProt) databases and more appropriately synchronize them in annotations of the protein-coding genes.

Proteome analyses of clinical samples with pertinent disease association have been of great resource for identification of “missing” proteins if high quality sequence data is searched against databases (Shiromizu et al. 2013). Further protein annotation information could be gained about PTMs using supplementary databases, such as PhosphoSitePlus (http://www.phosphosite.org). Bioinformatics annotations have been implemented searching for sequence similarities in standard databases selected for maximum functional characterization using a blast search strategy (Ranganathan et al. 2013).

It should be mentioned that new valuable data resources and data browsers, including Proteome Browser (Goode et al. 2013), CAPER (Guo et al. 2013) and CAPER2.0 (Wang et al. 2014), GenomewidePDB (Jeong et al. 2013) and Gene-centric Knowledgebase (Zgoda et al. 2013), emerge worldwide and are being developed to serve scientific community. There is a great need for databases organizing proteomic data in a gene-centric fashion, *e.g.*, GenomewidePDB, integrating proteomic (neXtProt) and transcriptomic (Ensembl) data with mass spectrometry (PeptideAtlas) and thus facilitating cross-identification of proteins at higher level. Specialized databases, such as the collection of nsSNP products used for identification of proteoform specific peptide sequences highlights that importance of the interplay between bioinformatic strategies and mass spectrometric quantifications (Song et al. 2014).

2.5 Application of Selected Reaction Monitoring Mass Spectrometry

Selected reaction monitoring (SRM) mass spectrometry is a protein sequencing methodology that utilizes the triple quadrupole tandem mass spectrometry technology for the quantification of proteins (Picotti et al. 2010; Lange et al. 2008). Recently, SRM applications in clinical studies represent a rapidly growing research area exploiting the high sequence specificity of transitions in SRM assays. In addition, the high sensitivity of modern MS instruments permits not only quantification but also targeted identification new proteoforms as HPP proposed (Paik et al. 2012b, 2014). Existing genomic databases form the base for selection of unique signature sequences of specific proteoforms, such as AST and nsSNP products. Following *in silico* digestion SRM assays can be developed specific for previously unidentified proteoforms using synthetic peptides. These peptides are then spiked into selected biological samples and used as internal standards to identify and further on, to quantify novel forms of proteins. The HPP projects take advantage of MS-based assays for at least two proteotypic peptides of any given protein and provide a reference set for the comprehensive quantitative coverage of the human proteome (Paik et al. 2012b).

As an example, a number of proteoforms of prostate specific antigen (PSA) were selected and SRM assays were developed based on curated and non-reviewed sequence variants in the UniProtKB database. The SRM assays were then applied on clinical samples attempting to find novel proteoforms of PSA identified at transcript level earlier. As a result, an nsSNP variant of PSA (L132I) was discovered that was never described at expression level before (Fig. 2.4) (Végvári et al. 2013). Due to the isobaric precursor and fragment ions of LSEPAELTDAVK and

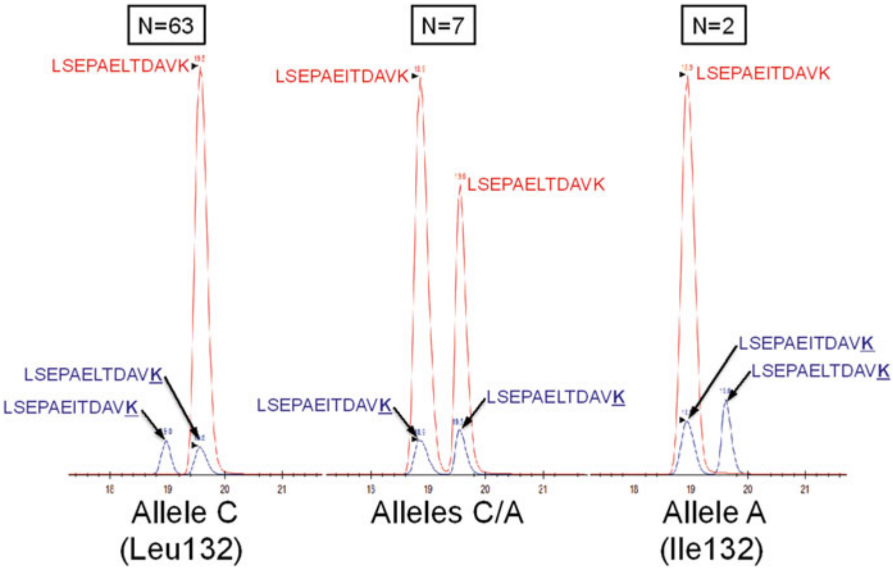


Fig. 2.4 Identification of a novel mutant PSA in clinical samples by SRM assay. The nsSNP (L132I) resulted in identical transitions of the peptides LSEPAELTDAVK and LSEPAEITDAVK that were chromatographically separated based on their hydrophobicity. Endogenous signals of LSEPAELTDAVK and LSEPAEITDAVK are shown in *red*, whereas their corresponding heavy labeled internal standard signals are in *blue* (Reproduced with permission from Végvári et al. 2013)

LSEPAEITDAVK peptides, identical transitions were observed the peaks were baseline separated by reversed-phase liquid chromatography. Since both proteoforms can be present in the same sample (heterozygous expression profile), the areas of both peaks have to be combined when quantifying the total amount of PSA. The population-based frequency of the allele A in exon 3 of the *KLK3* gene (dbSNP code: rs2003783) showed 10 % worldwide prevalence as reported in 1000 Genomes database.

The quantitative determination of PSA in blood has clinical importance in diagnosis and prognosis of prostate cancer (Lilja et al. 2007). However, the difficulties associated with clinical applicability of PSA, which can be at elevated concentration due to malignant as well as benign prostate disease, might be understood better if all molecular forms are precisely measured. Furthermore, the detailed insight to the microheterogeneity of PSA may reveal unknown diversity in the biology of prostate disease. The approach has more general interest since protein biomarkers are identified as differentially expressed in samples originating from disease and health clinical status. The cellular activity on all molecular forms of proteins in any certain disease state poses a high degree of complexity that has to be investigated in order to obtain clinically relevant insights.

2.6 Conclusions

It is a realistic view that the human proteome represents an enormous collection of sequences if all proteoforms, including up to an estimated one million different proteins derived by DNA recombination, alternative splicing of primary transcripts, point mutations and numerous post-translational modifications are considered. An additional assumption is that many of the modifications increase the number of functional proteoforms further by altering the primary products in a combinatorial manner.

However, the completion of the C-HPP project would already enhance understanding of human biology at cellular level and become a base for development of novel diagnostic, prognostic, therapeutic and preventive medical applications. In the future a large portion of the human proteome may be able determined is a single analysis provided that the development of mass spectrometric platforms used in proteome analyses continue to develop at today's pace.

References

Aebersold R, Bader GD, Edwards AM, et al. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. J Proteome Res. 2013;12(1):23–7.

Anderson NL, Anderson NG. The human plasma proteome – history, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1(11):845–67.

Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW. Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). J Proteome Res. 2004;3(2):235–44.

Berglund L, Bjorling E, Oksvold P, et al. A genecentric human protein atlas for expression profiles based on antibodies. Mol Cell Proteomic MCP. 2008;7(10):2019–27.

Björling E, Uhlén M. Antibodypedia, a portal for sharing antibody and antigen validation data. Mol Cell Proteomics. 2008;7(10):2028–37.

Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004;3(6):1234–42.

Desiere F, Deutsch EW, King NL, et al. The PeptideAtlas project. Nucleic Acids Res. 2006;34 Suppl 1:D655–8.

Domon B, Aebersold R. Review – mass spectrometry and protein analysis. Science. 2006;312(5771):212–7.

Fenyo D, Eriksson J, Beavis R. Mass spectrometric protein identification using the global proteome machine. Methods Mol Biol. 2010;673:189–202.

Gaudet P, Argoud-Puy G, Cusin I, et al. neXtProt: organizing protein knowledge in the context of Human Proteome Projects. J Proteome Res. 2013;12(1):293–8.

Goode RJ, Yu S, Kannan A, et al. The proteome browser web portal. J Proteome Res. 2013;12(1):172–8.

Guo F, Wang D, Liu Z, et al. CAPER: a chromosome-assembled human proteome browsER. J Proteome Res. 2013;12(1):179–86.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931–45.

Islam MT, Garg G, Hancock WS, Risk BA, Baker MS, Ranganathan S. Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the "Missing". J Proteome Res. 2014;13(1):76–83.

Jeong S-K, Lee H-J, Na K, et al. GenomewidePDB, a proteomic database exploring the comprehensive protein parts list and transcriptome landscape in human chromosomes. J Proteome Res. 2013;12(1):106–11.

Lane L, Argoud-Puy G, Britan A, et al. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res. 2012;40(Database issue):D76–83.

Lane L, Bairoch A, Beavis RC, et al. Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. J Proteome Res. 2014;13(1):15–20.

Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol. 2008;4:222.

Legrain P, Aebersold R, Archakov A, et al. The human proteome project: current state and future direction. Mol Cell Proteomic MCP. 2011;10(7):M111.009993.

Lilja H, Ulmert D, Bjork T, et al. Long-term prediction of prostate cancer up to 25 years before diagnosis of prostate cancer using prostate kallikreins measured at age 44 to 50 years. J Clin Oncol. 2007;25(4):431–6.

Lopez MF, Rezai T, Sarracino DA, et al. Selected reaction monitoring-mass spectrometric immunoassay responsive to parathyroid hormone and related variants. Clin Chem. 2010;56(2):281–90.

Mann M. Comparative analysis to guide quality improvements in proteomics. Nat Methods. 2009;6(10):717–9.

Martens L, Hermjakob H, Jones P, et al. PRIDE: the proteomics identifications database. Proteomics. 2005;5(13):3537–45.

Munoz J, Low TY, Kok YJ, et al. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol Syst Biol. 2011;7:550.

Olsen JV, Schwartz JC, Griep-Raming J, et al. A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. Mol Cell Proteomics. 2009;8(12):2759–69.

Paik YK, Jeong SK, Omenn GS, et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol. 2012a;30(3):221–3.

Paik YK, Omenn GS, Uhlen M, et al. Standard guidelines for the Chromosome-Centric Human Proteome Project. J Proteome Res. 2012b;11(4):2005–13.

Paik Y-K, Omenn GS, Thongboonkerd V, Marko-Varga G, Hancock WS. Genome-wide proteomics, chromosome-Centric Human Proteome Project (C-HPP), Part II. J Proteome Res. 2014;13(1):1–4.

Picotti P, Rinner O, Stallmach R, et al. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Methods. 2010;7(1):43–6.

Ranganathan S, Khan JM, Garg G, Baker MS. Functional annotation of the human chromosome 7 "Missing" proteins: a bioinformatics approach. J Proteome Res. 2013;12(6):2504–10.

Schmidt A, Claassen M, Aebersold R. Directed mass spectrometry: towards hypothesis-driven proteomics. Curr Opin Chem Biol. 2009;13(5–6):510–7.

Shi T, Fillmore TL, Sun X, et al. Antibody-free, targeted mass-spectrometric approach for quantification of proteins at low picogram per milliliter levels in human plasma/serum. Proc Natl Acad Sci U S A. 2012;109(38):15395–400.

Shiromizu T, Adachi J, Watanabe S, et al. Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-Centric Human Proteome Project. J Proteome Res. 2013;12(6):2414–21.

Smith BE, Hill JA, Gjukich MA, Andrews PC. Tranche distributed repository and ProteomeCommons.org. Methods Mol Biol. 2011;696:123–45.

Song C, Wang F, Cheng K, et al. Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. J Proteome Res. 2014;13(1):241–8.

The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 2011;39(Database issue):D214–9.

Végvári Á, Sjödin K, Rezeli M, et al. Identification of a novel proteoform of prostate specific antigen (SNP-L132I) in clinical samples by multiple reaction monitoring. Mol Cell Proteomic MCP. 2013;12(10):2761–73.

Wang D, Liu Z, Guo F, et al. CAPER 2.0: an interactive, configurable, and extensible workflow-based platform to analyze data sets from the Chromosome-Centric Human Proteome Project. J Proteome Res. 2014;13(1):99–106.

Whiteaker JR, Zhao L, Zhang HY, et al. Antibody-based enrichment of peptides on magnetic beads for mass-spectrometry-based quantification of serum biomarkers. Anal Biochem. 2007;362(1):44–54.

Whiteaker JR, Zhao L, Abbatiello SE, et al. Evaluation of large scale quantitative proteomic assay development using peptide affinity-based mass spectrometry. Mol Cell Proteomic MCP. 2011;10(4):M110.005645.

Zgoda VG, Kopylov AT, Tikhonova OV, et al. Chromosome 18 transcriptome profiling and targeted proteome mapping in depleted plasma, liver tissue and HepG2 Cells. J Proteome Res. 2013;12(1):123–34.

Zhong J, Cui Y, Guo J, et al. Resolving chromosome-centric human proteome with translating mRNA analysis: a strategic demonstration. J Proteome Res. 2014;13(1):50–9.

Chapter 3
Chromosome Transcriptome Profiling in the Context of High-Throughput Proteomics Studies

Elizabeth Guruceaga, Mariana B. Monteiro, María I. Mora, Lourdes Ortiz, Fernando J. Corrales, and Victor Segura

Abstract The Human Proteome Project was recently launched with the aim of mapping all protein species of the Human Proteome and to develop resources for a better understanding of human biology in health and disease. The complexity of the proteome and its cell/tissue specific distribution has led the research community to the integration of transcriptomics and proteomics landscapes in a proteogenomics approach. This strategy opens new research avenues that will surely enhance our capacity to annotate those gene products and proteoforms that still lack of experimental evidences In this chapter a bioinformatics pipeline to integrate RNA-Seq data, shotgun proteomics experiments and tissue specific gene expression patterns is proposed.

Keywords Transcriptome • Proteomics • Proteogenomics • C-HPP

3.1 Introduction

The sequencing of the human genome (Lander et al. 2001; Venter et al. 2001) has provided the first level of complexity of human biology. Despite this undoubted success, there is still a vast territory to be explored before a complete understanding of our own biology is achieved. A fundamental goal of genome projects is to

Fernando J. Corrales and Victor Segura contributed equally to this work.

E. Guruceaga (✉) • M.B. Monteiro • M.I. Mora • L. Ortiz • V. Segura
Genomics, Proteomics and Bioinformatics Unit, Center for Applied Medical Research,
University of Navarra, Pamplona, Spain
e-mail: eguruce@unav.es

F.J. Corrales
Genomics, Proteomics and Bioinformatics Unit, Center for Applied Medical Research,
University of Navarra, Pamplona, Spain

Division of Hepatology and Gene Therapy, Center for Applied Medical Research,
University of Navarra, Pamplona, Spain

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_3

generate a protein-coding catalog. The knowledge of the protein functions, regulatory mechanisms, networks of interaction, abundance and isoform patterns constitutes an essential issue for the understanding of human physiology in health and disease. However, unravelling the human proteome is a project that, despite the obvious a priori analogies with the sequencing of the human genome, represents a task that is far more challenging and whose boundaries still remain to be defined.

The Human Proteome Organization (HUPO) has coordinated the efforts of the international community promoting several initiatives to describe the human proteome in a systematic manner (http://www.hupo.org). One of these initiatives is the Human Proteome Project (HPP) (Legrain et al. 2011), designed in 2010 to map the entire human proteome using currently available and emerging techniques. The project is organised according to a chromosome-based strategy (C-HPP) where scientific groups from different nationalities agree to characterise the proteome of a selected chromosome (Paik et al. 2012a, b). All 24 chromosomes plus the mitochondrial genome encoded proteome have already been adopted by as many teams from 21 different countries. Even using state of the art technology, the proteomics community has encountered key challenges such as the low-abundance proteins and the complexity of protein isoforms present in a given cell. The vast heterogeneity, wide dynamic range and different ionisation efficiencies of proteins, among other reasons, restrict detection and quantification capacity on a large scale omics level (Nilsson et al. 2010).

C-HPP groups are now integrating transcriptomics data with proteomics, and relying on RNA sequencing methods to guide the genome-wide proteomics analyses (Paik and Hancock 2012). Combination of gene expression and proteomics information could suggest preferential cell lines where proteins might be detected and is currently an active working area in the biological annotation of chromosome 16 genes (Segura et al. 2013). In fact, this idea has been further developed by the Spanish HPP (SpHPP) consortium proposing a methodology to generate a transcriptomics map of chromosome 16 coding genes using RNA-Seq data from ENCODE project (ENCODE Project Consortium et al. 2012) and the Illumina Human Body Map (HBM). This approach falls within the scope of Proteogenomics and provides an effective mechanism for identifying cell lines or tissues with gene expression evidences for missing proteins (Segura et al. 2014).

Proteogenomics has recently emerged as a field in the junction of genomics and proteomics, although pioneer studies underlined the interest of such integrative approach (Nagaraj et al. 2011; Yates et al. 1995). The main goal is the matching of peptides identified in MS-based experiments against genome-wide gene/transcript sequence datasets for detailed gene annotation (Ansong et al. 2008), a method that has been successfully used to circumvent the limited availability of reference protein databases of non-model species (Evans et al. 2012). The integration of large amounts of RNA-Seq and MS data presents a challenging problem, starting from the generation of efficient and non-redundant RNA-Seq databases to search MS spectra (Woo et al. 2014). However, these difficulties have been successfully addressed to allow integration of high-throughput human proteome quantification with DNA variation and transcriptome information to reveal the multiple and diverse regulatory mechanisms of gene expression (Wu et al. 2013). Different critical points may benefit from crossing large-scale proteomics and transcriptomics/genomics datasets as proteogenomically identified peptides will provide unique information for gene annotation, such as confirmation of translation and prediction of novel genes (Castellana and Bafna 2010). Moreover, tissue/cell specific gene expression patterns will provide valuable information in the search of missing proteins and in the identification of protein variants resulting from alternative splicing, amino acid polymorphisms and post-translational modifications (Paik et al. 2012b).

3.2 Bioinformatics Methods in Proteogenomics

In this section, required resources for the proteome and genome annotation in the context of C-HPP are introduced. Then, bioinformatics and statistical methods used for the analysis and integration of transcriptomics and proteomics experiments are described.

3.2.1 Databases for Annotation of Genes and Proteins

The information about genes, transcripts and proteins and the relationships between the corresponding accession numbers have to be extracted from the following databases:

1. Ensembl (www.ensembl.org, release 73) provides the number of protein-coding genes.
2. neXtProt (www.nextprot.org, release 2013-10-10) provides the number of high-confidence proteins identified through mass spectrometry and other methods.
3. The Human Protein Atlas (www.proteinatlas.org, release 11.0) provides the number of proteins for which polyclonal antibodies have detected protein expression by immuno-histochemistry.

In the present phase of the C-HPP, research laboratories are focused in the detection of the missing proteins. neXtProt database has integrated peptide identification results extracted from Peptide Atlas (using 1 % FDR at protein level as threshold), bibliography, and direct submissions (Gaudet et al. 2013) and has defined five levels of protein evidence (PE):

1. Evidence at protein level: identification by mass spectrometry, detection by antibodies, Edman sequencing or resolved tridimensional structure
2. Evidence at transcript level: ESTs or full length mRNA expression
3. Inferred by homology: strong sequence similarity to known proteins in related species
4. Predicted: entries without evidences at PE levels 1, 2, 3 and 5

5. Uncertain: dubious sequences that are likely the products of erroneous translations of pseudogenes

The C-HPP missing proteins are established using the neXtProt protein existence evidences and are defined as those proteins with PEs from 2 to 4.

3.2.2 Analysis of Whole Transcriptome Experiments

The whole transcriptome can be studied with high throughput technologies such as expression microarrays or RNA-Seq. Data processing and statistical analyses for both approaches are summarised in Fig. 3.1. The processing of RNA-Seq samples is represented in blue in the figure, while all the steps corresponding to the data analysis of microarrays are shown in orange.

3.2.2.1 Microarray Data Analysis

In the case of microarray experiments there are well established protocols and methods for data analysis and normalisation (Fig. 3.1). Briefly, both background correction and normalisation is performed using the RMA (Robust Multichip Average) algorithm (Gentleman et al. 2004; Irizarry et al. 2003). R/Bioconductor (Gentleman et al. 2004) is used for preprocessing and statistical analysis. After normalisation, an expression threshold for each cell line is calculated to eliminate low intensity probe sets that can be considered technical noise. First, probe sets are sorted by their expression value in increasing order. For each probe set a *t*-test is performed to evaluate the differential expression between this probe set and the median value of the probe sets with lower expression levels. The *p*-values obtained are corrected for multiple hypothesis testing using FDR method (Storey and Tibshirani 2003), and FDR >0.95 (background signal) is considered as the criterion to calculate the corresponding intensity threshold (Segura et al. 2013). In the present chapter the mean value of the thresholds obtained for each sample (5.1 in logarithmic scale) is used to eliminate low intensity probe sets in all the samples.

3.2.2.2 Next-Generation Sequencing Data Analysis

Unlike the case of microarrays, methods for normalisation and statistical analysis of RNA-Seq data are not mature enough and no established best practices exist. However, it can be recommended as a guide the following pipeline (Fig. 3.1):

1. The downloaded sra files are converted into fastq files and the quality of the samples is verified using the open source software FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

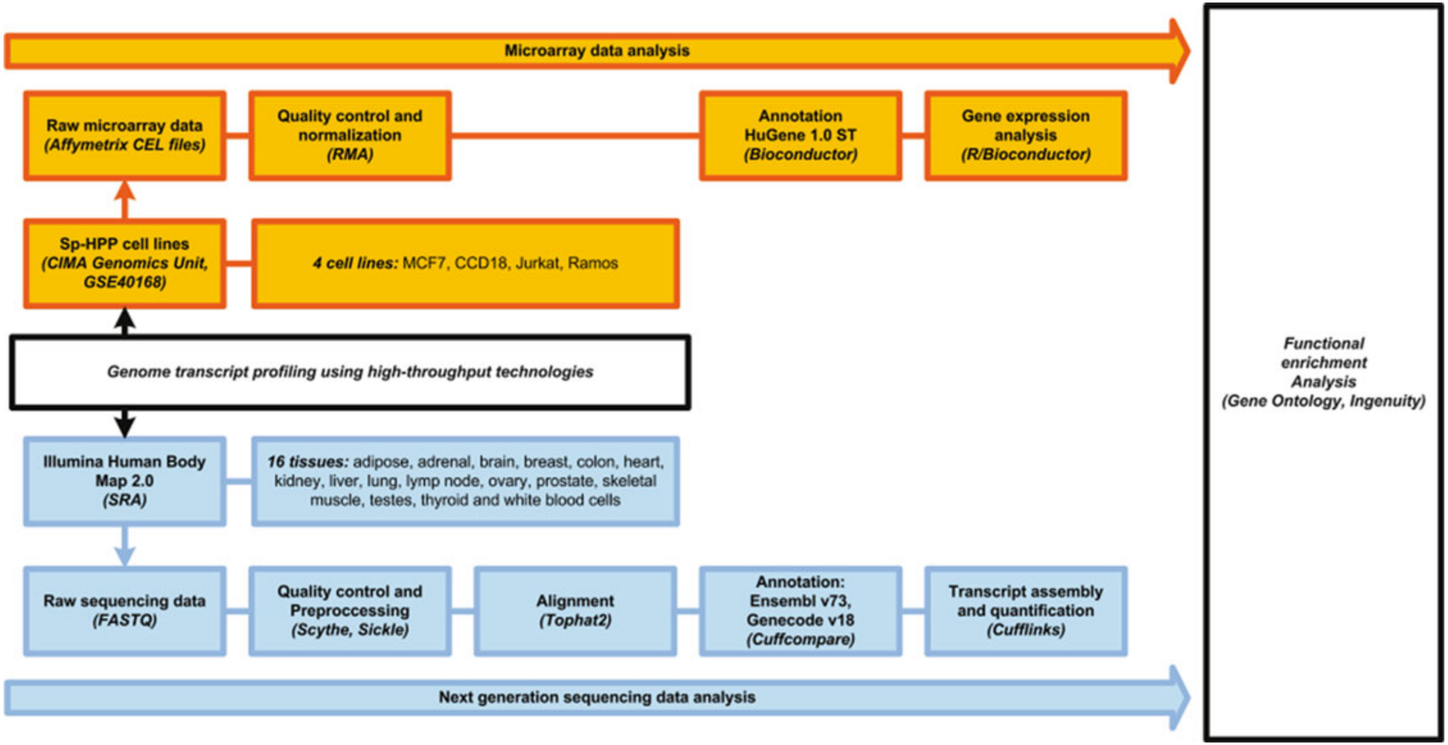

Microarray data analysis
Raw microarray data (Affymetrix CEL files)
Quality control and normalization (RMA)
Annotation HuGene 1.0 ST (Bioconductor)
Gene expression analysis (R/Bioconductor)
Sp-HPP cell lines (CIMA Genomics Unit, GSE40168)
4 cell lines: MCF7, CCD18, Jurkat, Ramos
Genome transcript profiling using high-throughput technologies
Illumina Human Body Map 2.0 (SRA)
16 tissues: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymp node, ovary, prostate, skeletal muscle, testes, thyroid and white blood cells
Raw sequencing data (FASTQ)
Quality control and Preproccessing (Scythe, Sickle)
Alignment (Tophat2)
Annotation: Ensembl v73, Genecode v18 (Cuffcompare)
Transcript assembly and quantification (Cufflinks)
Next generation sequencing data analysis
Functional enrichment Analysis (Gene Ontology, Ingenuity)

Fig. 3.1 Whole transcriptome analysis using microarray experiments (*orange*) and RNA-Seq data processing (*blue*)

2. The pre-processing of reads includes elimination of contaminant adapter substrings with Scythe (https://github.com/vsbuffalo/scythe) and quality-based trimming using Sickle (https://github.com/najoshi/sickle).
3. The alignment of reads to the human genome (hg19) is performed using Tophat2 mapper (Kim et al. 2013).
4. Transcript assembly and quantification using *Fragments Per Kilobase of transcript per Million fragments mapped* (FPKM) of genes and transcripts is carried out with Cufflinks2 (Trapnell et al. 2010).
5. The annotation of the obtained gene loci is performed using Cuffmerge with the corresponding reference genome.
6. Further analysis and graphical representations can be performed using the R/Bioconductor packages Biostrings (for the manipulation of biological sequences and files), doBy (for data processing) and ggplot2 (for the graphical data representation) (Gentleman et al. 2004).

In order to facilitate the interpretation of the expression analysis, the detected transcripts can be firstly divided into different categories based on the method published in (Ramsköld et al. 2009): not expressed genes (FPKM <0.3), lowly expressed genes (0.3 ≤ FPKM <3), medium expressed genes (3 ≤ FPKM <30), highly expressed genes (30 ≤ FPKM <100) and very highly expressed genes (FPKM ≥100). Moreover, genes expressed (FPKM ≥0.3) in all the studied samples are classified as ubiquitous genes while genes that are only expressed in a fraction of them are considered non-ubiquitous genes. Biological differences between these gene classes are found using enrichment analysis of Gene Ontology (GO) categories (G.O. Consortium 2013) with the hypergeometric test (Guruceaga et al. 2009) in R/Bioconductor.

3.2.3 Analysis of Whole Proteome Experiments

The data analysis procedure applied to mass spectrometry results has been previously described (Segura et al. 2014). This analysis starts with the translation of raw MS and MS/MS data to mascot general file format (mgf) and the search against UniProtKB/Swissprot database using Mascot (Matrix Science, London, UK). False discovery rate at protein level (FDR <1 %) has to be applied to obtained the protein identifications following HUPO guidelines, and protein grouping can be performed using PAnalyzer method (Prieto et al. 2012). For the integration of transcriptome and proteome results, Uniprot accession numbers obtained in the MS/MS queries need to be mapped into Ensembl gene identifiers with the bioinformatics tool PICR (Côté et al. 2007).

3.3 Chromosome Transcriptome Profiling Oriented Towards Proteome Research

The following sections contain the annotation of human proteome with neXtProt database. The achieved genome coverage of the protein coding genes is calculated and compared using a set of public microarrays and RNA-Seq experiments. The transcript profiling is also used to study the tissue specificity of gene expression profiles inferring the different biological functions in which are implicated ubiquitous and non-ubiquitous genes from enriched GO categories. Finally, shotgun proteomics results are combined with transcriptomics data to go in depth into the characterisation of missing proteins.

3.3.1 C-HPP Genome Annotation

The guidelines for the annotation of the chromosomes of human genome have been presented in the HUPO 12th World Congress held in Yokohama in 2013. The main source of information regarding proteins and genes is the neXtProt database (release 2013-10-10). The number of protein entries is 20,105, corresponding to 20,320 protein coding genes and 19,557 gene accession numbers from Ensembl release 73 (Table 3.1). The number of missing proteins is calculated as the number of protein coding genes with predicted protein evidence, homology evidence or transcript level evidence (PE2 + PE3 + PE4). The resulting number of missing proteins corresponds to 3,927 neXtProt entries. The chromosome distribution of protein evidences and missing proteins is presented in Table 3.1.

The combination of the human transcriptome and proteome data requires a common identifier for both experiments. The mapping efficiency of the Uniprot accession numbers obtained in MS/MS queries and Ensembl gene identifiers obtained in transcriptomics experiments into neXtProt identifiers must be as high as possible. Bioinformatics tool PICR (Côté et al. 2007) is recommended due to its good performance.

3.3.2 Whole Transcriptome Profiling Using Microarrays

In the present chapter a public microarray experiment has been analysed at genome level in order to calculate the theoretical neXtProt protein coverage that could be obtained with four human cell lines (Fig. 3.2a): Jurkat (human T cell lymphoblast-like cell line), CCD18 (human colon fibroblast cell line), MCF7 (human breast adenocarcinoma cell line) and Ramos (Human Burkitt's lymphoma cell line). This dataset is available to download in GEO database with the accession number GSE40168. The 88.41 % of the neXtProt entries have been detected at transcript

Table 3.1 Master table of protein evidences using neXtProt database release 2013-10-10, corresponding to Ensembl release 73

Chr	Nextprot entries	Protein coding genes	Uncertain (PE1)	Predicted (PE2)	Homology (PE3)	Transcript level (PE4)	Protein level (PE5)	Missing proteins
1	2,061	2,072	49	9	29	385	1,589	423
2	1,238	1,241	20	9	3	186	1,020	198
3	1,076	1,076	20	3	8	174	871	185
4	763	770	20	2	18	109	614	129
5	867	869	10	4	7	155	691	166
6	1,106	1,178	30	6	8	164	898	178
7	944	947	52	6	7	179	700	192
8	701	716	39	6	10	97	549	113
9	821	826	43	7	7	152	612	166
10	762	765	19	3	3	147	590	153
11	1,319	1,320	40	7	24	323	925	354
12	1,031	1,032	19	2	5	164	841	171
13	328	328	10	6	2	46	264	54
14	626	626	18	3	7	96	502	106
15	610	621	41	3	8	107	451	118
16	830	840	31	1	6	136	656	143
17	1,165	1,171	26	6	7	185	941	198
18	277	277	7	1	3	40	226	44
19	1,425	1,429	37	4	11	364	1,009	379
20	551	551	14	1	0	101	435	102
21	254	254	28	0	4	57	165	61
22	464	464	21	2	4	82	355	88
X	826	876	32	6	10	174	604	190
Y	46	57	8	0	0	16	22	16
MT	14	14	1	0	0	0	13	0
All	20,105	20,320	635	97	191	3,639	15,543	3,927

PE protein evidence

level in at least one of the four cell lines. It can be seen that the coverage along the chromosomes is rather uniform, without any of them presenting a remarkable low or high coverage (Fig. 3.2b).

Four public shotgun experiments (Segura et al. 2014) corresponding to the cell lines selected in (Segura et al. 2013) are also available in ProteomeXchange and the accession numbers are: PXD000443 (Jurkat), PXD000449 (CCD18), PXD000442 (MCF7) and PXD000447 (Ramos). Therefore, it is possible to evaluate the proteome coverage that can be obtained. We have identified a total of 8,433 proteins, 6,869 proteins detected in Jurkat cell line, 4,724 proteins in CCD18 cell line, 5,226 proteins in MCF7 cell line and 3,469 in Ramos cell line. It is interesting to highlight that the number of proteins identified in all the cell lines, 2,362 proteins, is compa-

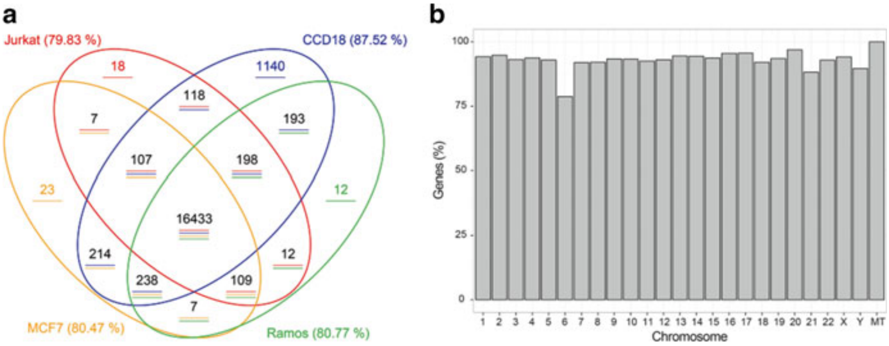


Fig. 3.2 Results of the microarray experiment in the Jurkat, CCD18, MCF7 and Ramos cell lines (GSE40168). (**a**) Venn diagram of the neXtProt coverage obtained with each microarray experiment. (**b**) Coverage of the neXtProt database obtained for each chromomose and measured as the percentage of protein coding genes detected in at least one of the microarray experiments

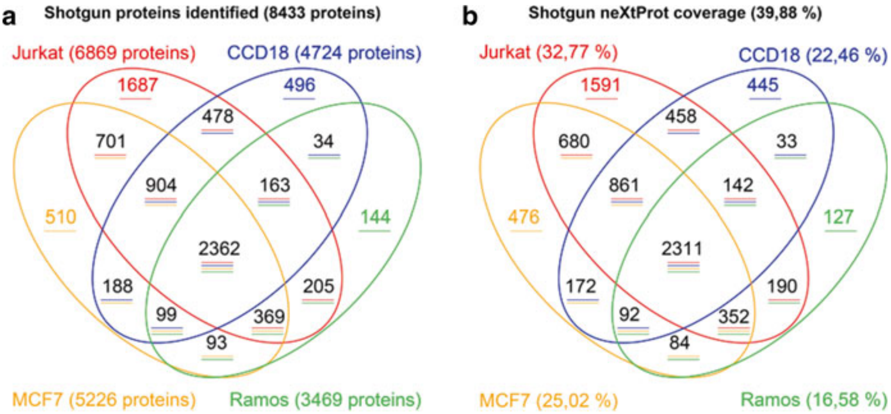


Fig. 3.3 Venn diagrams representing the results of the shotgun proteomics experiments in the Jurkat, CCD18, MCF7 and Ramos cell lines. (**a**) Proteins identified in the queries performed with MASCOT against the database UniprotKB/Swissprot. (**b**) Obtained coverage of the neXtProt database measured as the percentage of the protein coding genes detected in at least one of the shotgun experiments

rable and even smaller than the number of proteins detected only in one of the four studied cell lines, 2,837 proteins (Fig. 3.3a). This result suggests that the removal or inclusion of one cell line could alter significantly the total number of detected proteins. The neXtProt protein coding genes coverage obtained in the shotgun experiment is lower than expected given the transcriptome coverage for the same cell lines obtained in the microarray experiments. A total of 8,014 protein coding genes have been identified, which is the 39.88 % of the neXtProt entries (Fig. 3.3b). This means that more than half of the proteins coded by genes expressed in the microarrays have not been found in MS/MS experiments. In particular, 6,585 (32.77 %), 4,514 (22.46 %), 5,028 (25.02 %), and 3,331 protein coding genes (16.58 %) have been obtained in Jurkat, CCD18, MCF7 and Ramos cell lines respectively (Fig. 3.3b). As can be observed, the number of proteins detected in all the cell lines is 2,311, 28.84 % of all the proteins identified in the shotgun experiments, while as many as 2,639 proteins (32.93 %) have been found in only one of them.

This analysis can be considered as a proof-of-principle to show that the transcriptome profiling is a valuable source of information not only in Genomics research but also in a Proteomics context. However, there is still room for improvement in both the obtained transcriptome and proteome coverages. In the following section the obtained results for the studied cell lines have been enriched with the transcript profiling of 16 tissues using RNA-Seq, which is a more sensitive technology than the microarrays in the detection of low expressed genes (Ramsköld et al. 2009).

3.3.3 Extension of the Chromosome Transcriptome Profiling with RNA-Seq Data

The chromosome centric transcriptome landscape of human genome can be addressed based on next generation sequencing (NGS) technology. In particular, the public RNA-Seq experiments of 16 tissues included in the HBM dataset have been analysed in order to broaden the previously obtained transcriptome profiling based on microarray experiments.

The HBM project provides the transcriptome profiling of individual and mixture samples of 16 human tissues: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid and white blood cells. The accession numbers of these experiments are GSE30611 in GEO database, E-MTAB-513 in ArrayExpress and ERX011226 in SRA. The subset of samples of the HBM project that has been analysed in this chapter are detailed in Table 3.2.

The number of transcriptional units detected in the HBM is measured as the number of cufflinks genes. This number ranges from 39,970 in the liver sample to 160,710 in the lymph node sample. The mean number of gene assemblies in the 16 tissues is 88,578, including annotated and not annotated genes. The comparison of the transcript structures present in the HBM samples with the genome annotation available in Ensembl database (release 73) provides a total of 85,994 ensembl transcripts corresponding to 32,910 ensembl genes, 18,551 of which are protein coding genes. The minimum number of annotated transcripts is obtained in liver (21,254 ensembl transcripts) while the maximum number is found in lymph node (38,076 ensembl transcripts). The mean number of isoforms sequenced in the HBM samples using Ensembl annotation is 30,319 transcripts. In the case of genes the number varies between the 15,123 genes detected in liver and the 24,134 genes detected in testes sample, being the mean number of observed genes 19,245.

Table 3.2 List of samples of the Illumina Body Map 2.0 (HBM) dataset considered in the transcriptome analysis

SRA accession	GEO accession	Tissue	Library	Read count	Read mapped
ERX011215	GSM759490	Adipose	Single read	76269225	57712809
ERX011198	GSM759492	Adrenal	Single read	76171569	58552514
ERX011186	GSM759494	Brain	Single read	64313204	27810021
ERX011191	GSM759496	Breast	Single read	77195260	54787469
ERX011192	GSM759498	Colon	Single read	80257757	57472085
ERX011219	GSM759502	Kidney	Single read	79772393	59452594
ERX011183	GSM759500	Heart	Single read	76766862	40081383
ERX011211	GSM759504	Liver	Single read	77453877	38806672
ERX011222	GSM759506	Lung	Single read	81255438	64865353
ERX011188	GSM759508	Lymph node	Single read	81916460	58312329
ERX011214	GSM759512	Prostate	Single read	83319902	66255681
ERX011228	GSM759514	Skeletal muscle	Single read	82864636	64110513
ERX011208	GSM759520	White blood cells	Single read	82785673	55860468
ERX011196	GSM759510	Ovary	Single read	81003052	53470920
ERX011202	GSM759516	Testes	Single read	82044319	54447053
ERX011213	GSM759518	Thyroid	Single read	80246657	57993499

In an RNA-Seq analysis of gene expression, the quantification of transcripts is generally reported using FPKM as a measure of abundance. According to the thresholds previously calculated in (Ramsköld et al. 2009) for the definition of expression categories, the number of protein coding genes expressed in the HBM human tissues varies from 12,332 genes in liver, 57.90 % of the protein coding genes included in neXtProt database, to 16,688 genes in testes sample (78.35 %). The mean number of expressed genes per tissue is 14,414, corresponding to the 67.68 % of the neXtProt entries (Fig. 3.4a). The total percentage of neXtProt protein coding genes expressed in at least one of the tissues is 87.29 % (18,591 neXtProt entries). These results are in agreement with similar analysis of gene expression levels in human and mouse samples (Ramsköld et al. 2009). If each chromosome is considered independently, the mean percentage of genes detected per chromosome is 82.09 %. While the HBM dataset contains the 95.69 % of the chromosome 2 protein coding genes, none of the mitochondrial chromosome (MT) protein coding genes have been detected (Fig. 3.4b).

The comparison of Figs. 3.3 and 3.4 generates important findings related to the quantification of genes in tissues and cell lines. Surprisingly, the total coverage of neXtProt database is very similar (~87 %) and the individual coverage per cell line using microarrays is clearly better than the coverage per tissue using RNA-Seq. Moreover, the variance of the number of proteins per chromosome is also greater for RNA-Seq data, and significant differences are detected in certain chromosomes such as X, Y and MT. It is generally accepted that RNA-Seq is more sensitive, both in terms of detection of lowly expressed and differentially expressed genes (Ramsköld et al. 2009). However, the obtained results could be

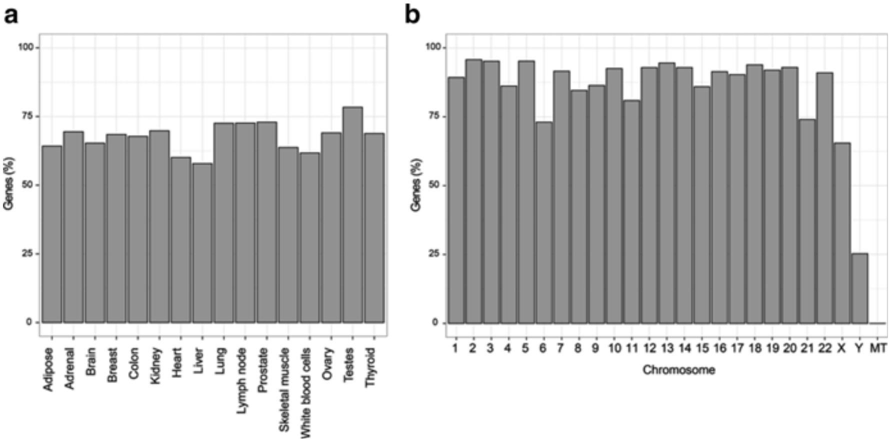


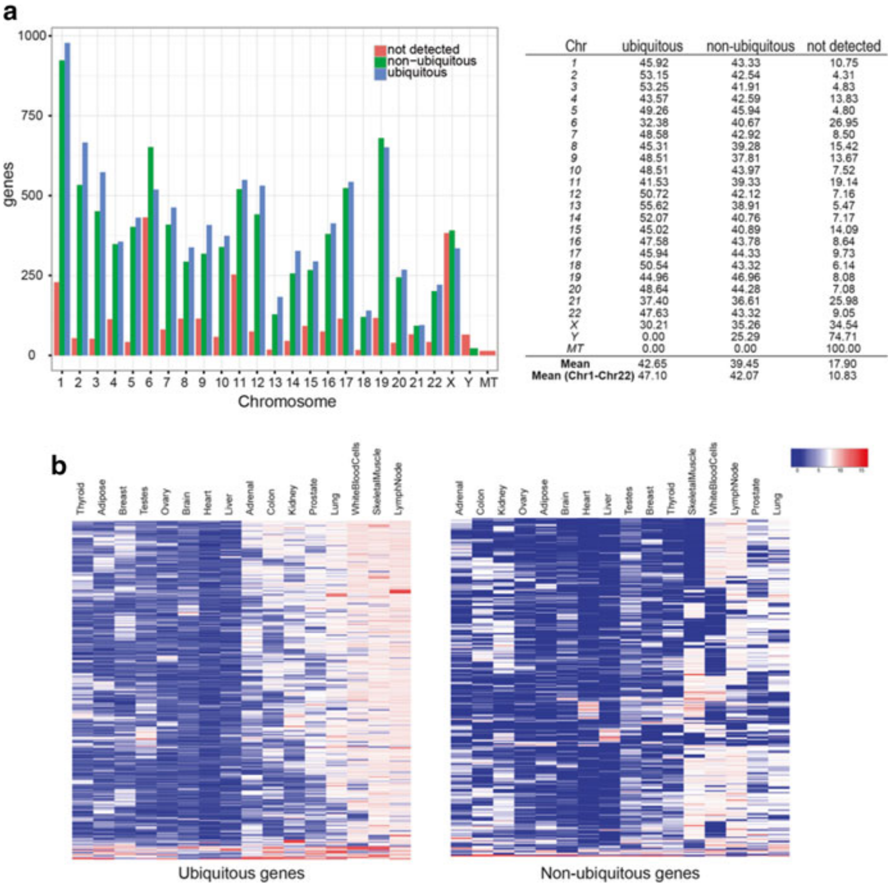
Fig. 3.4 Distribution of the number of protein coding genes detected in the HBM dataset. (**a**) Distribution of the detection percentage along tissues. (**b**) Distribution of the detection percentage along chromosomes

explained by a quantification of expression levels in RNA-Seq data highly dependent on read depth. A low read coverage inevitably reduces the number of genes whose expression can be assessed and reduces the power to detect differences in expression with RNA-Seq. At the same time a probe-specific background hybridization on the microarrays has been previously described for genes with large microarray intensities but small sequence counts in RNA-Seq (Marioni et al. 2008; Black et al. 2014).

3.3.3.1 Tissue Specific Expression of Human Genes

The study of gene expression patterns in the HBM facilitates the characterisation of genes with different levels of expression and with distinct tissue specificity. These gene profiles could be related to specific biological processes or functions.

Genes have been classified into three categories based on their tissue specificity: (1) *not detected genes*, genes not expressed in any of the 16 tissues; (2) *ubiquitous genes*, genes that are expressed in all the 16 tissues at any expression level; (3) *non-ubiquitous genes*, genes that are expressed only in one or in a group of the analysed tissues. As a result of this classification, 2,707 not detected genes in HBM (12.71 % of the neXtProt protein coding genes), 9,656 ubiquitous genes (45.34 %) and 8,935 non-ubiquitous genes (41.95 %) have been found. The gene distribution of these specificity categories in each chromosome is shown in Fig. 3.5a. The mean number of not detected (~20 %), ubiquitous (~40 %) and non-ubiquitous (~40 %) genes remains constant for all the chromosomes except X, Y and MT chromosomes.



Chr	ubiquitous	non-ubiquitous	not detected
1	45.92	43.33	10.75
2	53.15	42.54	4.31
3	53.25	41.91	4.83
4	43.57	42.59	13.83
5	49.26	45.94	4.80
6	32.38	40.67	26.95
7	48.58	42.92	8.50
8	45.31	39.28	15.42
9	48.51	37.81	13.67
10	48.51	43.97	7.52
11	41.53	39.33	19.14
12	50.72	42.12	7.16
13	55.62	38.91	5.47
14	52.07	40.76	7.17
15	45.02	40.89	14.09
16	47.58	43.78	8.64
17	45.94	44.33	9.73
18	50.54	43.32	6.14
19	44.96	46.96	8.08
20	48.64	44.28	7.08
21	37.40	36.61	25.98
22	47.63	43.32	9.05
X	30.21	35.26	34.54
Y	0.00	25.29	74.71
MT	0.00	0.00	100.00
Mean	42.65	39.45	17.90
Mean (Chr1-Chr22)	47.10	42.07	10.83

Fig. 3.5 (**a**) Number of protein coding genes of each category (*left*) and their percentage (*right*) in each chromosome. (**b**) Expression profiles of a set of ubiquitous and non-ubiquitous genes (*Blue* represents low expression levels and *red* represents high expression levels)

Clustering analysis of ubiquitous and non-ubiquitous genes has been performed to visualise the differences in their expression profiles. For this purpose, the 500 genes of each category with a greater variance across the 16 tissues are previously selected (Fig. 3.5b). Interestingly, the obtained heatmaps show tissue specific profiles for non-ubiquitous genes while the expression profiles of ubiquitous genes are very similar to each other.

Finally, a functional enrichment analysis based on GO categories (G.O. Consortium 2013) has been performed to evaluate the possible differences in the function, cellular localisation and biological processes in which are involved the ubiquitous and non-ubiquitous genes. Applying the hypergeometric distribution (Guruceaga et al. 2009) with a *p-value* <0.01 threshold, several enriched categories have been found. In Fig. 3.6 the 10 GO terms with best *p-values* are represented for each gene

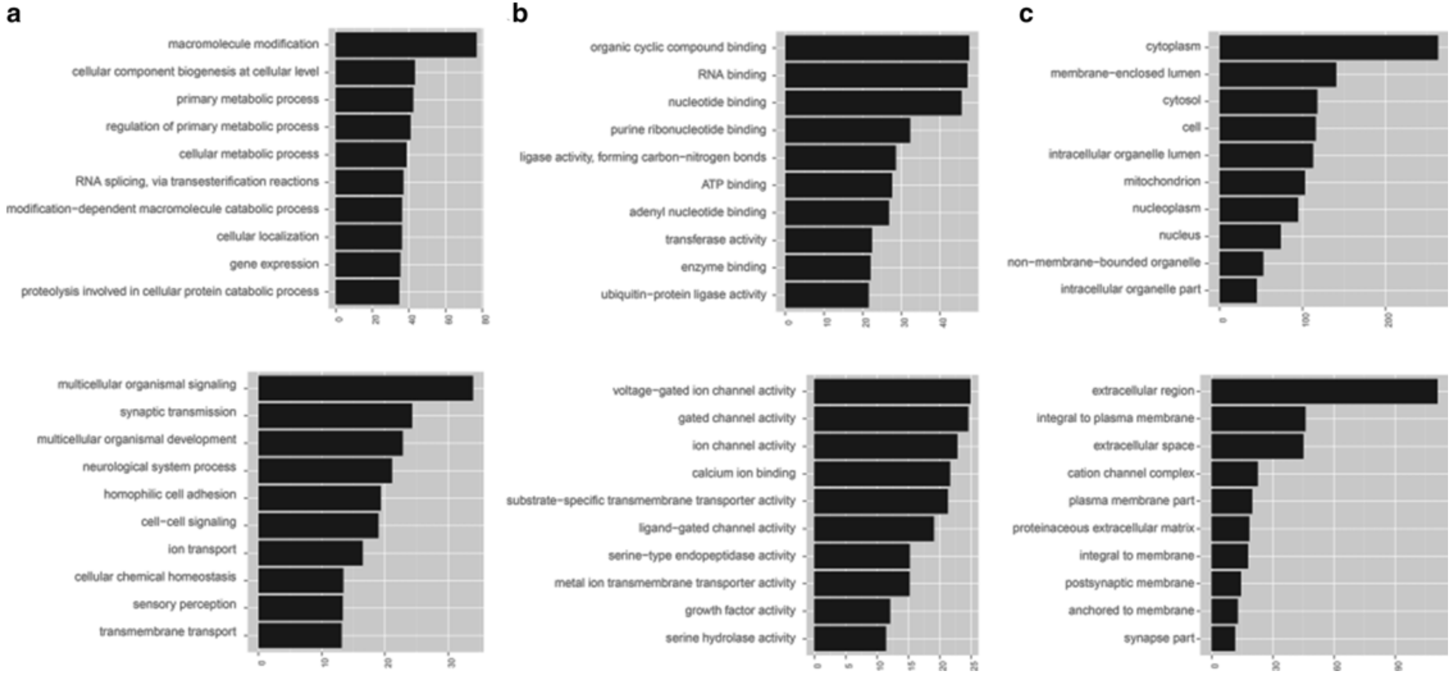


Fig. 3.6 Graphical representation of the gene ontology enrichment analysis for each of the three ontologies: (**a**) Biological process. (**b**) Molecular function. (**c**) Cellular component. Only the ten most significant categories are included in each of the *bar* plots and the height of the *bars* corresponds to their (−10)*log(p-value); (*top*) results for the list of ubiquitous genes, (*bottom*) Results for non-ubiquitous genes

specificity category and each ontology (BP, Biological Process; MF, Molecular Function; CC, Cellular Component). The preferred localisations of non-ubiquitous genes are the cell membrane and the extracellular region, while ubiquitous genes appear inside the cell (cytosol, mitochondrion and nucleus). On the other hand, ubiquitous genes enrich core biological processes such as metabolic functions, control of gene expression or macromolecule activity, as opposed to the more specific functions (signaling, transporter activity or synaptic transmission) mediated by non-ubiquitous genes. These results are in agreement with previous observations about this issue (Ramsköld et al. 2009).

The results of the analysis of gene specific profiles in HBM tissues support the idea of carefully selecting the tissue or cell line where the detection of certain proteins is going to be attempted in order to improve the number of detections.

3.3.3.2 In-Depth Analysis of Missing Proteins

In the context of the HPP project special attention should be devoted to the characterisation and detection of the missing proteins. These proteins share a lack of any experimental evidence in the neXtProt database. The transcriptome profiling obtained with RNA-Seq and microarray experiments and the integration of these data with the proteomics shotgun experiments results could be of great help for addressing this challenge.

The first question that can be investigated is if significant expression differences exist between the missing and the non-missing, or known, proteins according to the HBM. The FPKM distributions for both gene sets in each analysed tissue are represented in Fig. 3.7a. Applying a *t*-test between the obtained distributions it is concluded that there is a statistically significant decrease in the expression of the protein coding genes for missing proteins in all tissues except skeletal muscle ($p < 0.01$). Assuming RNA-protein correlation, this result suggests that missing proteins are less abundant, a possible justification for the difficulties encountered in their experimental detection.

The total number of missing proteins in the human proteome is 3,923 neXtProt entries, corresponding to 4,135 protein coding genes. The mean number of missing protein coding genes for each chromosome is 165 (Fig. 3.7b). The mitochondrial chromosome (MT) does not include missing protein coding genes, while chromosome 1 has the highest number of them (445 genes). The transcriptome profiling previously generated using high-throughput technologies can be used as a guide to infer the expression level of the missing proteins in cell lines (microarrays) and tissues (HBM). The expression of 3,315 missing protein coding genes have been detected (84.50 % of neXtProt missing proteins) in the microarray experiments of the cell lines (Fig. 3.8a), whereas in the HBM samples 2,940 (74.94 %) are considered expressed in at least one of the tissues (Fig. 3.8c). The classification of gene tissue specificity may give us a clue about the localisation of the missing proteins. The enrichment analysis of the set of missing protein genes in the not detected

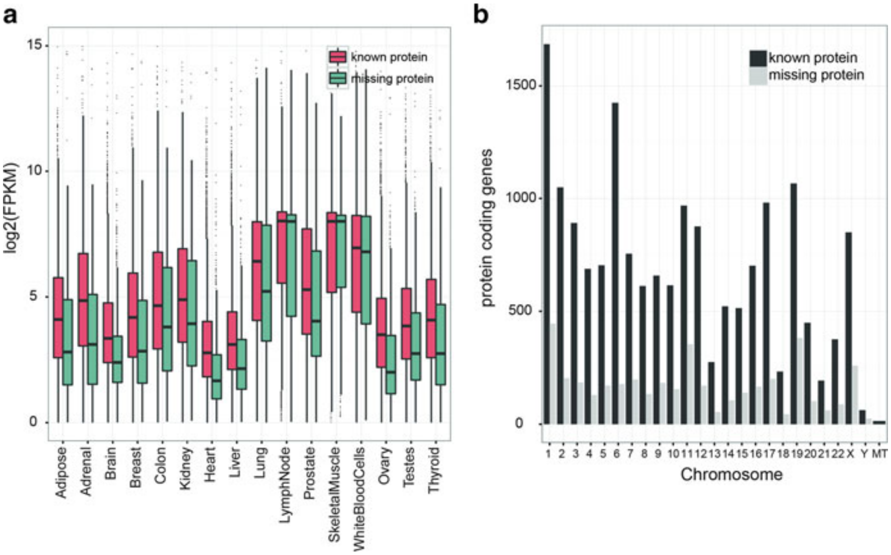


Fig. 3.7 Remarkable results of the analysis of missing proteins. (**a**) FPKM distribution in the 16 tissue samples of HBM for missing and known protein coding genes. (**b**) Distribution of the number of missing protein coding genes across chromosomes

(1,167 genes), ubiquitous (497 genes) and non-ubiquitous (2,471 genes) gene sets reveals a statistical significance enrichment in the former ones ($p < 1e-12$).

Considering the proteins identified in the shotgun experiments, 147 missing proteins (3.75 % of the missing proteins in the human genome) have MS/MS evidence. In Jurkat cell line 61 proteins have been found, in CCD18 cell line 56, in MCF7 cell line 58 and finally 25 proteins have been detected in Ramos cell line (Fig. 3.8b). The comparison of the information provided by transcriptomics and proteomics experiments requires careful analysis and interpretation (Fig. 3.8c). The majority of proteins identified in shotgun experiments have been seen in the genomics analysis (113 proteins), while the corresponding genes of 7 of these proteins are not expressed in any cell line or analysed tissues. Despite the higher number of genes detected in the microarray experiments, using the same cell lines of the proteomics experiments, only 16 identified proteins are specific to these samples. Moreover, 11 identified proteins are coded by genes only present in HBM tissues. These results are in agreement with the previous observation of a higher correlation between protein expression and RNA-Seq measured gene expression than that obtained with microarray gene expression (Fu et al. 2009).

The expression levels of the missing protein coding genes in the RNA-Seq data can be used to rank the tissues based on the number of these genes that are expressed in each tissue. Lymph nodes, skeletal muscle, testes and white blood cells result to be the preferred biological samples for new proteomics experiments (Fig. 3.8d).

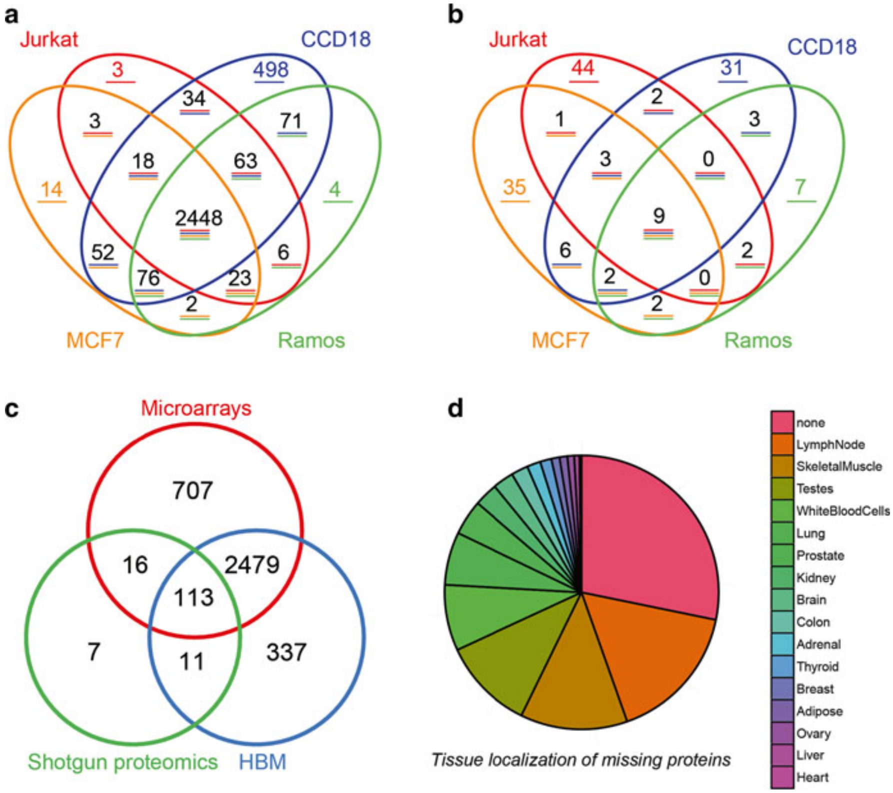


Fig. 3.8 (**a**) Venn diagram of missing proteins detected in the microarray experiments of cell lines. (**b**) Venn diagram of missing proteins detected in shotgun experiments performed in spHPP. (**c**) Venn diagram with the summary of transcriptomics and proteomics experimental evidences for missing proteins. (**d**) Pie chart of the number of missing protein coding genes expressed in each tissue

Finally, the study of missing proteins concludes with a clustering of their corresponding gene expression profiles in the HBM (Fig. 3.9a) and a functional enrichment analysis in order to determine the localisation and biological processes regulated by them. The clustering analysis reveals the more probable sample in which a particular protein can be detected. The functional analysis points out a biological similarity between the non-ubiquitous genes and missing proteins (Fig. 3.9b). This observation is supported by a statistically significant enrichment of missing proteins in the gene set of non-ubiquitous genes.

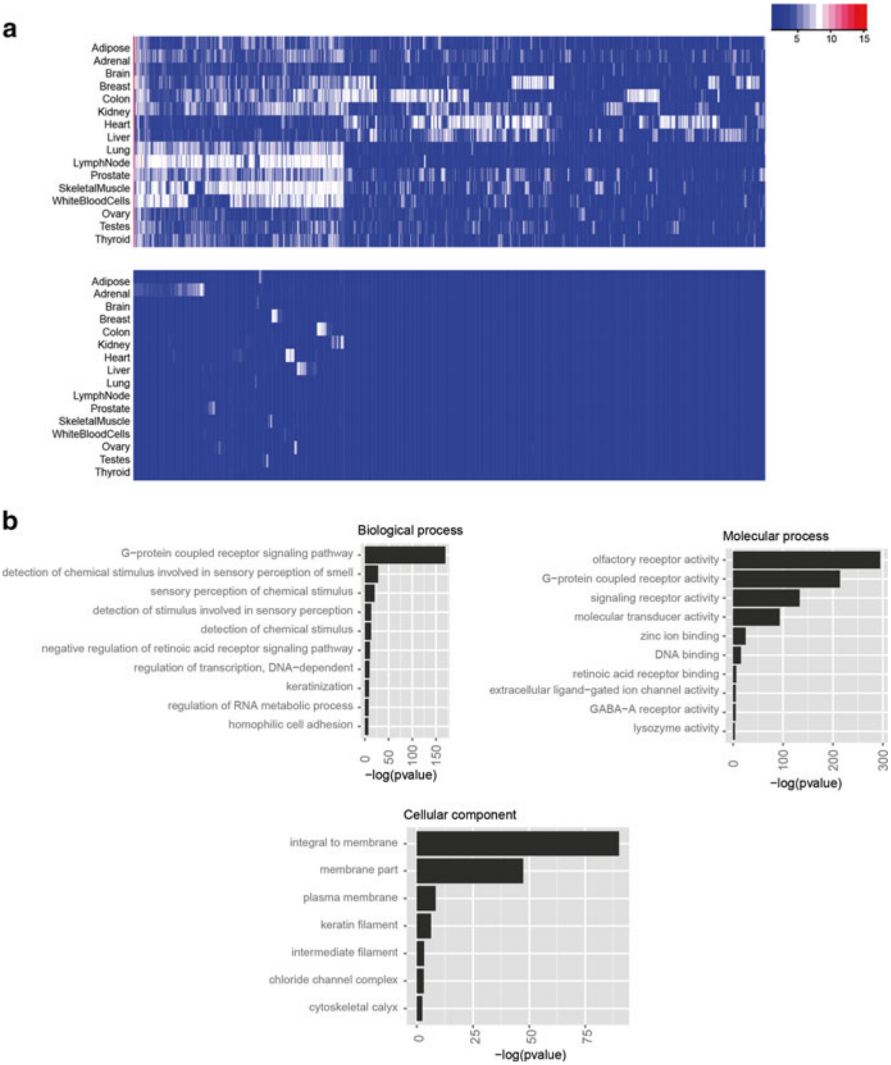


Fig. 3.9 Clustering (**a**) and functional enrichment analysis (**b**) of the missing protein genes

3.4 Summary

In this chapter, detailed transcript profiling of protein coding genes is described within the context of C-HPP. The Human Proteome Project (HPP) has defined several guidelines to characterise all the proteins encoded by the human genome. The experimental strategy followed by the international research groups which are part of the C-HPP initiative includes proteomics shotgun experiments in order to detect proteins in a diversity of tissues and cell lines. One of the most difficult tasks in this project is the detection of proteins without any previous experimental evidence. neXtProt is the reference database of experimental protein evidences using spectrometry-based identifications or antibody assays from which these proteins without any evidence, named missing proteins, are extracted.

In previous studies, genes of chromosome 16 were analysed using microarrays for the CCD18, MCF7, Jurkat and Ramos cell lines (Segura et al. 2013) and RNA-Seq data for 16 tissues of the HBM (Segura et al. 2014). This information has been enriched extending the analysis of HBM RNA-Seq data and deepening the interpretation of the results. Genes have been classified based on their expression levels and their presence in one, several or all the studied tissues. This approach has been applied not only to the whole genome but also at the chromosome level. A functional analysis of the defined gene categories has shown biological differences between the obtained classes. In particular, this analysis has allowed establishing gene expression profiles and functions related to the ubiquitous and non-ubiquitous set of genes. These results confirm the tissue specificity of certain transcripts that eventually could be responsible of a lower detection probability of their corresponding proteins. Therefore, the study of the transcriptome using high-throughput technologies (microarrays, RNA-Seq) can be extremely useful for the proper selection of the biological sample where the missing proteome of each chromosome could be detected.

Other important sources of information are the public shotgun experiments performed by the spHPP. The integration of the proteome obtained in these cell lines (Jurkat, MCF7, CCD18 and Ramos) with the gene quantification of HBM tissues generates meaningful results about the success in the characterisation of missing proteins. A more detailed study of these proteins in the transcriptome has prioritised the best tissues in which new experiments should be carried out, including lymph nodes, skeletal muscle, testes and white blood cells. Despite these endeavors, there remain many challenges ahead of us. Roughly 28 % of the set of the missing proteins lack of any transcriptome evidence in the HBM tissues, so a new set of samples with other tissues and cell lines must be analysed using the proposed workflow.

Acknowledgments All participating laboratories are members of ProteoRed-ISCIII. This work was supported by: ProteoRed and the Carlos III National Health Institute Agreement, ProteoRed-ISCIII; the agreement between FIMA and the “UTE project CIMA”; grants SAF2011-29312 from Ministerio de Ciencia e Innovación and ISCIII-RETIC RD06/0020 to FJC. We also thank the Proteomics, Genomics and Bioinformatics Core Facility of CIMA.

References

Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic. 2008;7(1):50–62.

Black MB, Parks BB, Pluta L, Chu TM, Allen BC, Wolfinger RD, Thomas RS. Comparison of microarrays and RNA-seq for gene expression analyses of dose-response experiments. Toxicol Sci. 2014;137(2):385–403.

Castellana N, Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics. 2010;73(11):2124–35.

Côté RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, Leinonen R, Apweiler R, Hermjakob H. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. BMC Bioinform. 2007;8:401.

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunte C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods. 2012;9(12):1207–11.

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics. 2009;10:161.

G.O. Consortium. Gene ontology annotations and resources. Nucleic Acids Res. 2013;41:D530–5.

Gaudet P, Argoud-Puy G, Cusin I, Duek P, Evalet O, Gateau A, Gleizes A, Pereira M, Zahn-Zabal M, Zwahlen C, Bairoch A, Lane L. neXtProt: organizing protein knowledge in the context of human proteome projects. J Proteome Res. 2013;12(1):293–8.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

Guruceaga E, Segura V, Corrales FJ, Rubio A. FactorY, a bioinformatic resource for genome-wide promoter analysis. Comput Biol Med. 2009;39(4):385–7.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31(4):e15.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach

H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, Bergeron J, Borchers CH, Corthals GL, Costello CE, Deutsch EW, Domon B, Hancock W, He F, Hochstrasser D, Marko-Varga G, Salekdeh GH, Sechi S, Snyder M, Srivastava S, Uhlén M, Wu CH, Yamamoto T, Paik YK, Omenn GS. The human proteome project: current state and future direction. Mol Cell Proteomics. 2011;10(7):M111.009993.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.

Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Pääbo S, Mann M. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol. 2011;7:548.

Nilsson T, Mann M, Aebersold R, Yates 3rd JR, Bairoch A, Bergeron JJ. Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods. 2010;7(9):681–5.

Paik YK, Hancock WS. Uniting ENCODE with genome-wide proteomics. Nat Biotechnol. 2012;30(11):1065–7.

Paik YK, Omenn GS, Uhlen M, Hanash S, Marko-Varga G, Aebersold R, Bairoch A, Yamamoto T, Legrain P, Lee HJ, Na K, Jeong SK, He F, Binz PA, Nishimura T, Keown P, Baker MS, Yoo JS, Garin J, Archakov A, Bergeron J, Salekdeh GH, Hancock WS. Standard guidelines for the chromosome-centric human proteome project. J Proteome Res. 2012a;11(4):2005–13.

Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee HJ, Na K, Choi EY, Yan F, Zhang F, Zhang Y, Snyder M, Cheng Y, Chen R, Marko-Varga G, Deutsch EW, Kim H, Kwon JY, Aebersold R, Bairoch A, Taylor AD, Kim KY, Lee EY, Hochstrasser D, Legrain P, Hancock WS. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol. 2012b;30(3):221–3.

Prieto G, Aloria K, Osinalde N, Fullaondo A, Arizmendi JM, Matthiesen R. PAnalyzer: a software tool for protein inference in shotgun proteomics. BMC Bioinform. 2012;13:288.

Ramsköld D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol. 2009;5(12):e1000598.

Segura V, Medina-Aunon JA, Guruceaga E, Gharbi SI, González-Tejedo C, Sánchez del Pino MM, Canals F, Fuentes M, Casal JI, Martínez-Bartolomé S, Elortza F, Mato JM, Arizmendi JM, Abian J, Oliveira E, Gil C, Vivanco F, Blanco F, Albar JP, Corrales FJ. Spanish human proteome project: dissection of chromosome 16. J Proteome Res. 2013;12(1):112–22.

Segura V, Medina-Aunon A, Mora MI, Martínez-Bartolomé S, Abian J, Aloria K, Antúnez O, Arizmendi JM, Azkargorta M, Barceló-Batllori S, Beaskoetxea J, Bech-Serra JJ, Blanco F, Monteiro MB, Cáceres D, Canals F, Carrascal M, Casal JI, Clemente F, Colomé N, Dasilva N, Díaz P, Elortza F, Fernández-Puente P, Fuentes M, Gallardo O, Gharbi SI, Gil C, González-Tejedo C, Hernáez ML, Lombardía M, Lopez-Lucendo M, Marcilla M, Mato JM, Mendes ML, Oliveira E, Orera I, Pascual-Montano A, Prieto G, Ruiz-Romero C, Sánchez Del Pino MM, Tabas-Madrid D, Valero ML, Vialas V, Villanueva J, Albar JP, Corrales FJ. Surfing transcriptomic landscapes. A step beyond the annotation of Chromosome 16 proteome. J Proteome Res. 2014;13(1):158–72.

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440–5.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Publ Group. 2010;28:511–5.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science. 2001;292(5523):1838.

Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, MacCoss M, Bafna V. Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res. 2014;13(1):21–8.

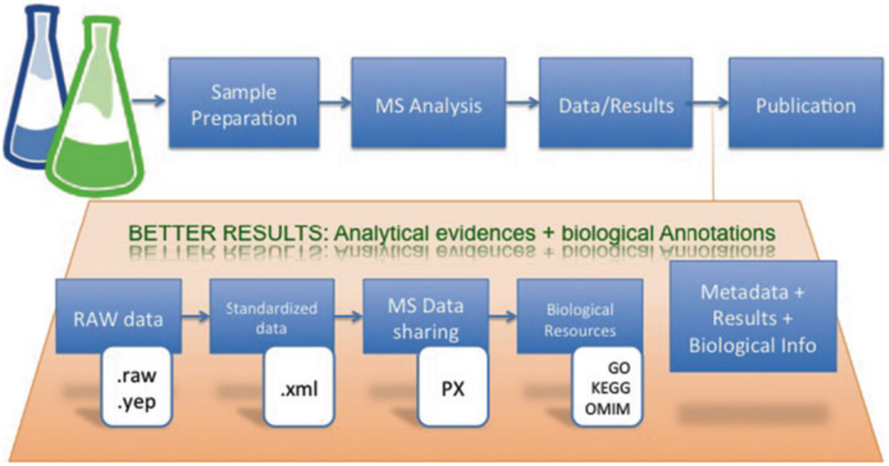
Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, Tang H, Snyder M. Variation and genetic control of protein abundance in humans. Nature. 2013;499(7456):79–82.

Yates 3rd JR, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem. 1995;67(18):3202–10.

Chapter 4
Standards for Proteomics Data Dissemination and Experiments Reporting

J. Alberto Medina-Aunon and Juan P. Albar

Graphical Abstract



Sample extraction, separation and data acquisition/processing are sequentially executed to get protein expression profiles from a given sample. To retrieve all the information embedded in the result files, it is important to translate raw files into accepted standards (MIAPE (Minimum Information About a Proteomics Experiment) guidelines and HUPO-PSI xml formats). Data standardization enables to normalize the data obtained from distinct experimental platforms. Remarkably, it provides a way to reanalyse datasets post-publication using public repositories such as ProteomeXchange (PX). Thanks to these repositories, scientists

J.A. Medina-Aunon • J.P. Albar (✉)
Spanish National Center for Biotechnology-CSIC, Madrid, Spain

Carlos III Networked Proteomics Platform, ProteoRed-ISCIII, Madrid, Spain
e-mail: jpalbar@proteored.org

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_4

are able to download and reanalyse the data using specific search-criteria or with up-to-date databases, to refine or complement their studies.

Finally, proteomics results can be linked to functional and biological databases to integrate gathered information in a systems biology network. Several resources regarding metabolic pathways (KEGG, Biocarta), their reported molecular interactions (IntAct, STRING), their possible role in diseases (OMIM), as well as their posttranslational modifications (Phosphosite, Phosida), biological functions or molecular process (Gene Ontology, GO) can all be compiled through curated and freely available repositories.

Keywords Standardization • Proteomics • Genomics • Databases • Protein sequences

4.1 Background: Basic Principles

Nowadays, thousands of proteins can be identified and quantified in proteomic experiments. However, for proteomics to reach its technological maturity and to be able to successfully deliver all its potential to the scientific community, there is a need for quality control (QC) procedures that cover all the steps involved in the proteomics analyses. It is crucial that the large amount of data that is being gathered is of high reliability, and importantly that it can be shared among the scientific community in easily interchangeable formats.

Several efforts to develop test standards and QC procedures have been undertaken by different organizations such as the Association of Biomolecular Resource Facilities (ABRF), the Human Proteome Organization (HUPO) or the National Cancer Institute (NCI) and the Spanish Network of Proteomics Facilities (ProteoRed). All these have underscored the importance of standardization and QC to improve the quality of the data and to evaluate both the robustness and reproducibility of proteomics workflows. Standardization in the way to report proteomic experiments is being defined through several guidelines which have been established by the community over the recent years all grouped under the "Minimal Information About a Proteomic Experiment" (MIAPE). The definition of standard exchangeable data formats has also been the subject of a long-term effort followed by the Proteomics Standard Initiative of the Human Proteome Organization (HUPO-PSI).

Implementing correct annotation of the experimental proteomics reports will enable the reproduction and re-analysis of the results, allowing in this way to achieve one of the fundamental principles of the scientific research. However, the historic heterogeneity of data formats released by equipment vendors has hindered this process and made it more challenging to verify the reported results even in published studies. All these issues are of utmost importance and consequently open other topics like to define the type of repositories where proteomics data should be shared, exchanged, downloaded or even re-processed. Early works from outstanding proteomics groups and consortia have aimed to define public repositories for proteomics data sets with agreed formats to store and exchange the analytical information derived from proteomics laboratories (Taylor et al. 2003; Carr et al. 2004; Garwood et al. 2004; Pedrioli et al. 2004).

The proteomics data generated has to be inserted in or linked to other pre-existing framework of knowledge. In this sense, proteomics data deposited in major proteomics repositories must be integrated with the information gathered from genome wide repositories. Understanding the functional biology depends on the integration of all molecular building blocks that are being unravelled in a systems biology strategy. To achieve this process, we need to know the main features of the array of genomics and proteomics databases where the functional knowledge is deposited, as well as the available bioinformatics tools used to integrate them. This current review addresses some of these issues, although not in a fully comprehensive way.

4.2 Data Standards: HUPO-PSI-XML, MIAPEs and CVS

Data standards and standardized use of terminologies and ontologies for biomedical informatics are critical to report high-throughput experimental results in formats that can be interpreted by researchers or analytical tools. This is essential in the "Omics-science" field.

The initiative of standardization within the Human Proteome Organisation, HUPO-PSI, has a large experience (Orchard et al. 2012) in improving data standardisation, drawing together academic and industrial partners in an open collaboration. The stated aims of the HUPO-PSI are to "define community standards for data representation in proteomics to facilitate data comparison, exchange and verification" (Kaiser 2002; Orchard et al. 2003). From its foundation in 2002, this group has held annual meetings, workshops and published community documents that have contributed significantly to this objective.

The main organizational units of the HUPO-PSI are the workgroups devoted to specific proteomics areas, namely Molecular Interactions (MI) Mass Spectrometry (MS), Proteomics Informatics (PI), and Protein Separation (PS), (http://www.psidev.info/) Each workgroup has the following document outputs (Fig. 4.1):

1. Guidelines for reporting the Minimum Information About a Proteomics Experiment (MIAPE reports). These are mainly focussed on the experiment metadata for allowing its replication based on the original conditions.
2. Formal exchange formats usually represented in Extensible Markup Language (XML) for communicating data among software packages or sending results to

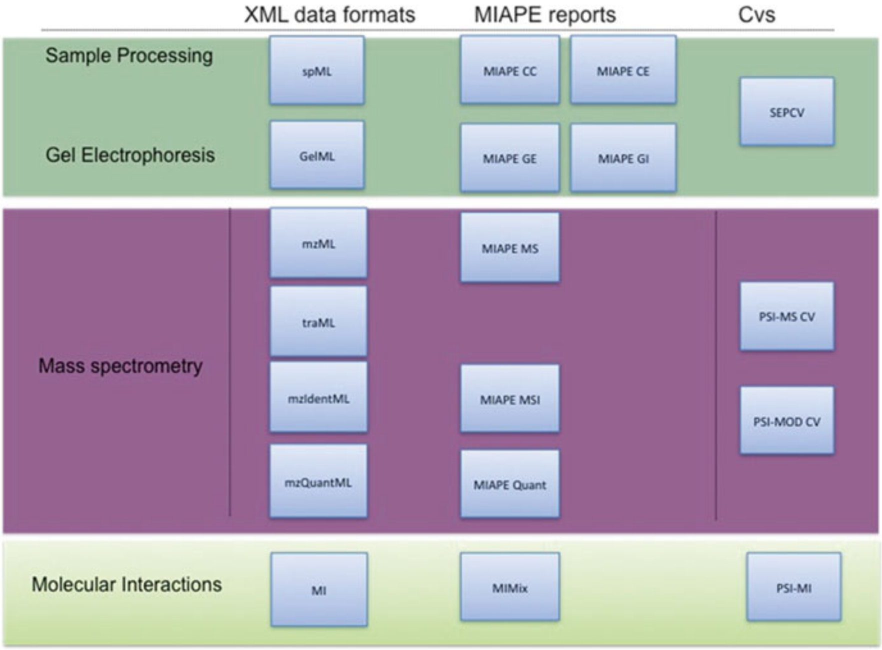


Fig. 4.1 Summary of HUPO PSI milestones. XML Exchange formats, MIAPE guidelines and controlled vocabularies

a public databases. They normally capture experimental results as well as significant additional details relative to the experimental workflow performed.

3. Syntactic and semantic validation by controlled vocabularies (CVs) to provide a standard terminology for the data elements represented within data formats.

4.2.1 Reporting Guidelines: Minimal Information About a Proteomic Experiment (MIAPE Documents)

In 2007, the HUPO-PSI published a collective report to define the Minimum Information About a Proteomics Experiment (MIAPE) specifications (Taylor et al. 2007). From this root document, a set of MIAPE documents for the different proteomics workflow stages were also delivered (Taylor et al. 2008; Binz et al. 2008; Gibson et al. 2008; Hoogland et al. 2010; Jones et al. 2010; Domann et al. 2010; Martinez-Bartolome et al. 2013). Each MIAPE module collects a minimal checklist of items that should be reported for a given technique associated to each stage. The items were stated using plain language and they describe specific points related not only to the experimental protocol descriptors but also to the data analysis that has been performed. These descriptors, usually called metadata should guide other groups to interpret the published results without ambiguity as well as to know how they were generated.

A first MIAPE document, the MIAPE-MS, comprises the essential elements that correspond to the mass spectrometry analysis. This MIAPE–MS (Taylor et al. 2008) is divided in specific sections that report the mass spectrometer configuration (ion source, analyzer and detector) and describe the method used to generate the resulting peak list.

A second document, the MIAPE-MSI (Binz et al. 2008) focuses on the mass spectrometry-based peptide and protein identification process and the steps that characterize the proteomics experiment. The information related to the input data (MS data), the search parameters (database, tolerances, endoprotease, etc…) and the results, including identified proteins and peptides, as well as their interpretation and validation, should be included in this report.

Other documents devoted to protein and peptide separation techniques such as the MIAPE guidelines for gel-based experiments (MIAPE-GE (Gibson et al. 2008)), gel-based image analysis (MIAPE-GI (Hoogland et al. 2010)), column chromatography MIAPE-CC (Jones et al. 2010) and capillary electrophoresis MIAPE-CE (Domann et al. 2010) have also been published.

Finally, the most recent MIAPE document, MIAPE-Quant (Martinez-Bartolome et al. 2013), describes a wide range of MS-based quantitative approaches, including peptide or protein labelling, targeted-MS and label-free approaches. Fundamentally, a MIAPE-Quant report contains the experimental design, the description of the starting biological material and the conditions used to perform the work. This includes sample description, assays descriptors, group structure (resulting from experimental conditions and their values) and replicate structure. Additionally, the reference to the dataset used for quantitative analysis, the quantitative protocol and the resulting data could also be included.

4.2.2 The Extensible Markup Language (XML) to Standardize Proteomics Data Storage

In parallel, the HUPO-PSI has also developed data exchange formats (Martens et al. 2011; Deutsch et al. 2012; Jones et al. 2012; Walzer et al. 2013; Gibson et al. 2010) applicable to each experimental module, making data sharing and reporting less time-consuming for both proteomics experts and occasional users. In the context of a typical proteomics Liquid Chromatography (LC)-MS based workflow, the initial step at the level of mass spectrometry information is to integrate unprocessed and processed spectra (e.g. peak lists for peptide/protein identification). For this purpose the HUPO-PSI XML data standard is called mzML (Martens et al. 2011). This format allocates the relevant information related to the MS acquisition step in a basic XML architecture. On the other hand, the standardized format mzIdentML (Jones et al. 2012) contains the output of peptide/protein identification that is generated by the database search engine.

In addition, for identification (and quantitation) by selected reaction monitoring analyses (SRM), the TraML standard format (Deutsch et al. 2012) captures the list of compounds, the retention time or the input acquisition parameters used to generate the transition.

The HUPO-PSI has recently released two other standard formats for capturing the output from quantitation software the mzQuantML (Walzer et al. 2013), an XML format capturing a detailed evidence trail, and the mzTab (Griss et al. 2014), a tabular-separated format capturing a simple summary of results.

In addition, the HUPO-PSI has also released the GelML standard format (Gibson et al. 2010) to describe gel-based experiments and results. Finally, the HUPO-PSI has also played a very active role in the standardisation of protein interaction data since almost a decade ago (Orchard et al. 2007), although work on protein interactions is outside of the scope of this chapter.

4.2.3 Controlled Vocabularies

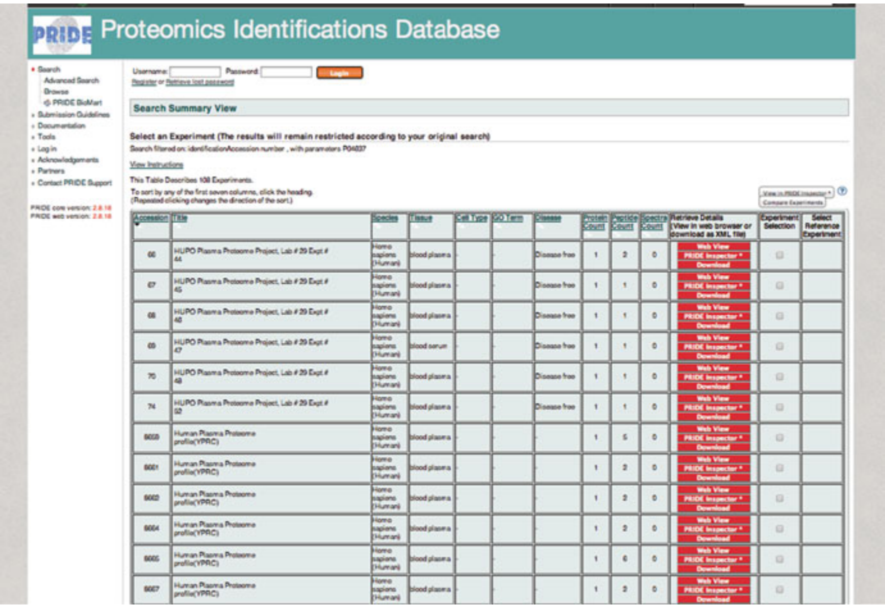
HUPO-PSI's Controlled Vocabularies (CVs) provide a consensus annotation system to standardize the meaning, syntax and formalism of terms used across proteomics experiments. Each PSI working group has developed the CVs required by the technology or data type and it aims to standardize, following common recommendations for development and maintenance. In addition, it is proposed a common mapping schema to describe for each exchange schema the associations between its specific elements and the rest of PSI CVs or other external ontology resources.

4.3 Repositories for Proteomics and Biological Data

Data publication is the final stage of most of the experimental pipelines. The submission of both the results and metadata to public repositories enables further reanalysis of the data that eventually may complement the initial studies. Here, we firstly describe some of the most popular public repositories regarding the proteomics experimental data and secondly, the main databanks where it is possible to retrieve functional, biological and pathological data linked to a single protein or a set of them. Both elements will bring a comprehensive explanation of proteomics studies, providing at the same time the analytical evidences and the biological background of the identified proteins.

4.3.1 MS-Based Data: ProteomeXchange Consortium

The ProteomeXchange consortium has been set up to provide a coordinated submission of MS proteomics data to the main existing proteomics repositories: PRIDE, PeptideAtlas and GPMdb. In the following paragraphs, a short description of any of



PRIDE core version: 2.8.18
PRIDE web version: 2.8.18

Search Summary View

Select an Experiment (The results will remain restricted according to your original search)

Search filtered on: identificationAccession number , with parameters P04637

View Instructions

This Table Describes 108 Experiments.

To sort by any of the first seven columns, click the heading.
(Repeated clicking changes the direction of the sort.)

Accession	Title	Species	Tissue	Cell Type	GO Term	Disease	Protein Count	Peptide Count	Spectra Count	Retrieve Details (View in web browser or download as XML file)	Experiment Selection	Select Reference Experiment
66	HUPO Plasma Proteome Project, Lab # 29 Expt # 44	Homo sapiens (Human)	blood plasma	-	-	Disease free	1	2	0	Web View PRIDE Inspector * Download	□	
67	HUPO Plasma Proteome Project, Lab # 29 Expt # 45	Homo sapiens (Human)	blood plasma	-	-	Disease free	1	1	0	Web View PRIDE Inspector * Download	□	
68	HUPO Plasma Proteome Project, Lab # 29 Expt # 46	Homo sapiens (Human)	blood plasma	-	-	Disease free	1	1	0	Web View PRIDE Inspector * Download	□	
69	HUPO Plasma Proteome Project, Lab # 29 Expt # 47	Homo sapiens (Human)	blood serum	-	-	Disease free	1	1	0	Web View PRIDE Inspector * Download	□	
70	HUPO Plasma Proteome Project, Lab # 29 Expt # 48	Homo sapiens (Human)	blood plasma	-	-	Disease free	1	1	0	Web View PRIDE Inspector * Download	□	
74	HUPO Plasma Proteome Project, Lab # 29 Expt # 52	Homo sapiens (Human)	blood plasma	-	-	Disease free	1	1	0	Web View PRIDE Inspector * Download	□	
8660	Human Plasma Proteome profile(YPRC)	Homo sapiens (Human)	blood plasma	-	-	-	1	5	0	Web View PRIDE Inspector * Download	□	
8661	Human Plasma Proteome profile(YPRC)	Homo sapiens (Human)	blood plasma	-	-	-	1	2	0	Web View PRIDE Inspector * Download	□	
8662	Human Plasma Proteome profile(YPRC)	Homo sapiens (Human)	blood plasma	-	-	-	1	2	0	Web View PRIDE Inspector * Download	□	
8664	Human Plasma Proteome profile(YPRC)	Homo sapiens (Human)	blood plasma	-	-	-	1	2	0	Web View PRIDE Inspector * Download	□	
8665	Human Plasma Proteome profile(YPRC)	Homo sapiens (Human)	blood plasma	-	-	-	1	6	0	Web View PRIDE Inspector * Download	□	
8667	Human Plasma Proteome profile(YPRC)	Homo sapiens (Human)	blood plasma	-	-	-	1	2	0	Web View PRIDE Inspector * Download	□	

Fig. 4.2 PRIDE Query. List of public projects where the protein P04637 (Cellular tumour antigen p53) was detected. Just introducing the UniProtKB accession it is possible to browse the experimental details regarding its identification by mass spectrometry

them is provided. The idea behind ProteomeXchange is to provide a single point for data submission and deposition to afterwards, allow multiple points of access for data visualization and analysis. A complete list of the public experiments could be accessed and downloaded at http://proteomecentral.proteomexchange.org/.

While this chapter was writing, ProteomeXchange defined two pipelines for data submission depending of the nature of the data to upload. Thus MS/MS data is processed and checked using PRIDE based pipeline (http://www.proteomexchange.org/submission), meanwhile SRM/MRM (Selected/Multiple Reaction Monitoring) data is processed through PASSEL (PeptideAtlas SRM Experiment Library) (Farrah et al. 2012) framework (http://www.peptideatlas.org/passel/). Both ways offer a straightforward and easy way to upload the data and detailed user guides.

4.3.1.1 PRIDE (http://www.ebi.ac.uk/pride/)

The PRotein IDEntification Database (PRIDE) (Vizcaino et al. 2013) is a centralized public data repository focussed on protein and peptide identifications developed by the EMBL-EBI (Hinxton, UK). PRIDE can store mass spectra (peak lists), peptide and protein identifications together as much their associated metadata as possible (Fig. 4.2). In spite of PRIDE was developed prior to the establishment of the current stable standards mzML and mzIdentML, an XML schema (PRIDE XML) was defined as an essential complementary tool to this main public repository. It describes how data should be represented and uploaded to PRIDE. The widely use of this XML-based format has been assumed as data standard for reporting MS and MSI data within a single file. When a MS-based dataset is uploaded to ProteomeXchange, automatically it is published at the same time in PRIDE. In fact, currently this is the only way to publish a dataset in PRIDE.

By January 2014, 1495 projects, containing 31665 MS-based experiments were stored within this database. In terms of proteins, peptides and mass spectra, the last reference in 2012 (Vizcaino et al. 2013) included around 11.1 million identified proteins 61.9 million identified peptides and 324 million spectra.

4.3.1.2 PeptideAtlas (http://www.peptideatlas.org/)

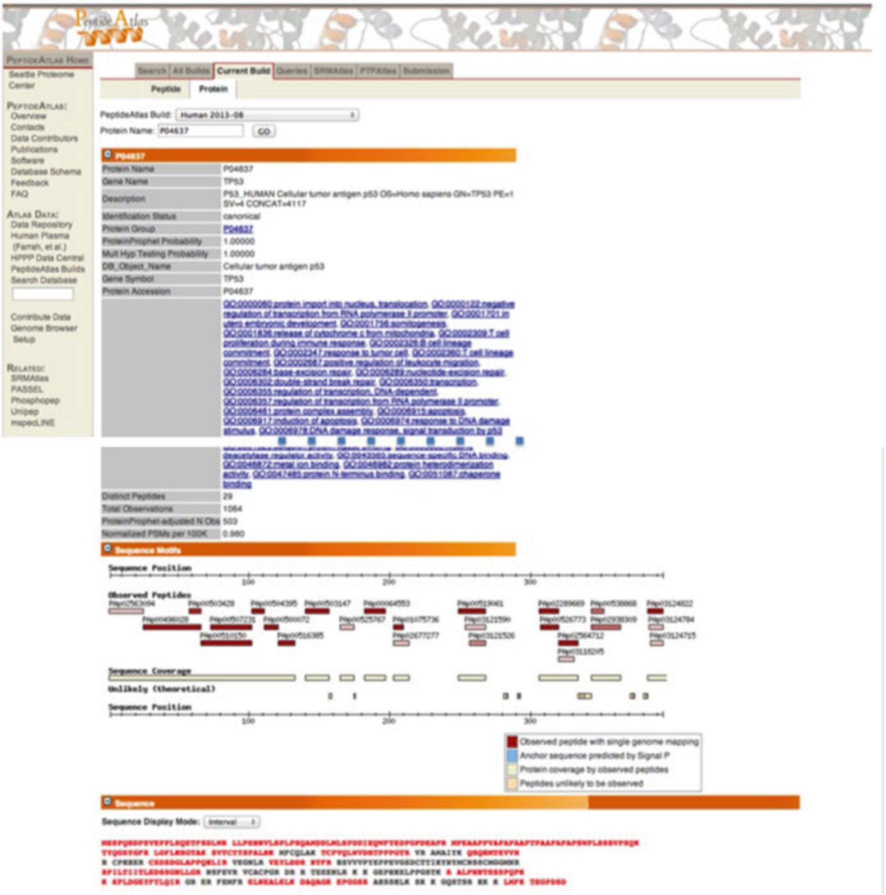
PeptideAtlas developed by the Institute for Systems Biology (Seattle, USA). It is a multi-organism, publicly accessible compendium of identified peptides in large sets of tandem mass spectrometry proteomics experiments (Deutsch 2010; Desiere et al. 2006) (Fig. 4.3). Mass spectra files for human, mouse, yeast, and several other organisms were collected and searched using the latest search engines and protein sequences, including a reprocessing of the data through the Trans Proteomic Pipeline (Deutsch et al. 2010). It brings low false discovery rates, thus providing an accurate and up to date vision of a particular proteome. The identified peptide sequences are then mapped onto their respective genome sequence, resulting into species or sample specific "builds", which represent all peptides mapped to a single reference. By January 2014, more than 1,000 experiments, containing the raw data and the search results were available. About data statistics, the last reference in 2010 (Deutsch 2010) pointed out around 90,000,000 spectra, 10,000,000 peptides and 42,500 distinct proteins.

In addition to the traditional MS repository, the PeptideAtlas also supports selected/multiple reaction monitoring (SRM/MRM) proteomics experiments. This complementary repository, SRMAtlas (http://www.srmatlas.org), is a compendium of targeted proteomics assays to detect and quantify proteins in complex proteome digests by MS. It results from high-quality measurements of natural and synthetic peptides conducted on a triple quadrupole mass spectrometer.

An extensive form is available at https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_Submit to upload both targeted (SRM/MRM) and discovery (shotgun) data. In the case of SRM data, results (transition list) may be formatted following the HUPO-PSI TraML schema.

4.3.1.3 Global Protein Machine Database – GPM/gpmDB (http://www.thegpm.org/)

Originally developed by Beavis Informatics Ltd (Canada), the Global Proteome Machine Database (gpmDB) is the proteomics hub with the highest amount of MS/MS data, including spectra, peptide and protein identifications (Fig. 4.4). By January



PeptideAtlas Build: Human 2013-08

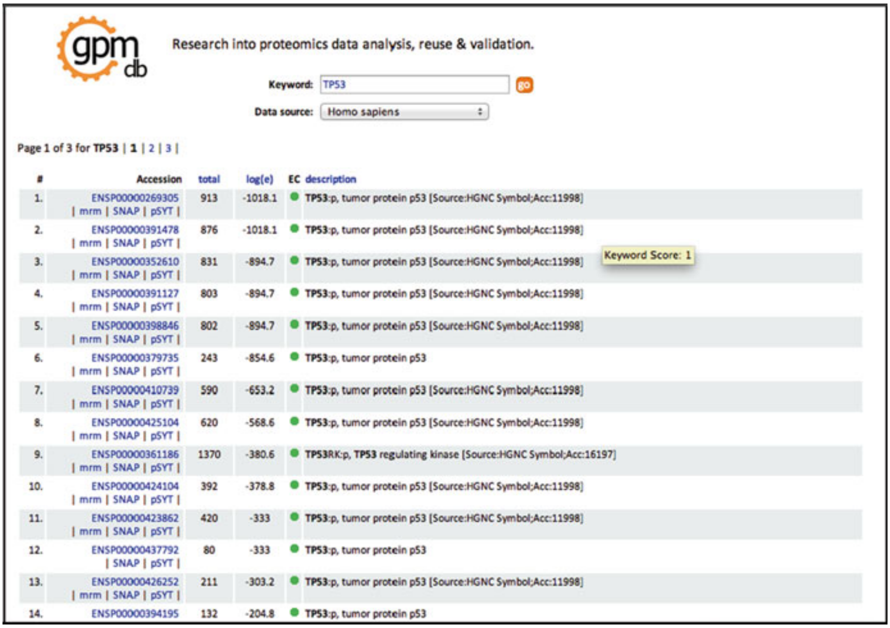
Protein Name: P04637

P04637

Protein Name	P04637
Gene Name	TP53
Description	P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens GN=TP53 PE=1 SV=4 CONCAT=4117
Identification Status	canonical
Protein Group	P04637
ProteinProphet Probability	1.00000
Mult Hyp Testing Probability	1.00000
DB_Object_Name	Cellular tumor antigen p53
Gene Symbol	TP53
Protein Accession	P04637
	GO:0000060 protein import into nucleus, translocation, GO:0000122 negative regulation of transcription from RNA polymerase II promoter, GO:0001701 in utero embryonic development, GO:0001756 somitogenesis, GO:0001836 release of cytochrome c from mitochondria, GO:0002309 T cell proliferation during immune response, GO:0002326 B cell lineage commitment, GO:0002347 response to tumor cell, GO:0002360 T cell lineage commitment, GO:0002687 positive regulation of leukocyte migration, GO:0006284 base-excision repair, GO:0006289 nucleotide-excision repair, GO:0006302 double-strand break repair, GO:0006350 transcription, GO:0006355 regulation of transcription, DNA-dependent, GO:0006357 regulation of transcription from RNA polymerase II promoter, GO:0006461 protein complex assembly, GO:0006915 apoptosis, GO:0006917 induction of apoptosis, GO:0006974 response to DNA damage stimulus, GO:0006978 DNA damage response, signal transduction by p53
	deacetylase regulator activity, GO:0043565 sequence-specific DNA binding, GO:0046872 metal ion binding, GO:0046982 protein heterodimerization activity, GO:0047485 protein N-terminus binding, GO:0051087 chaperone binding
Distinct Peptides	29
Total Observations	1064
ProteinProphet-adjusted N Obs	503
Normalized PSMs per 100K	0.980

Fig. 4.3 PeptideAtlas query: available information for P04637 protein. Among other data, it is possible to display the different identified peptides by mass spectrometry

2014, it counted 251,000 models (experiments), and more than 148,840,000 and 1,183,000,000 proteins and peptides respectively (http://wiki.thegpm.org/wiki/Main_Page). This database has been integrated into the GPM server pages, allowing users to quickly compare their experimental results with the best results that have been previously stored in the database. MS/MS data is processed mostly using the popular open source search engine X!Tandem (Craig and Beavis 2004). Peptide and protein identifications are generated and stored in the gpmDB. One of the ways to access the data is through the built proteomes (collection of taxonomy-specific divisions).



Research into proteomics data analysis, reuse & validation.

Keyword: TP53

Data source: Homo sapiens

Page 1 of 3 for **TP53** | **1** | 2 | 3 |

<table>
<tr><th>#</th><th>Accession</th><th>total</th><th>log(e)</th><th>EC</th><th>description</th></tr>
<tr><td>1.</td><td>ENSP00000269305
| mrm | SNAP | pSYT |</td><td>913</td><td>-1018.1</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>2.</td><td>ENSP00000391478
| mrm | SNAP | pSYT |</td><td>876</td><td>-1018.1</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>3.</td><td>ENSP00000352610
| mrm | SNAP | pSYT |</td><td>831</td><td>-894.7</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>4.</td><td>ENSP00000391127
| mrm | SNAP | pSYT |</td><td>803</td><td>-894.7</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>5.</td><td>ENSP00000398846
| mrm | SNAP | pSYT |</td><td>802</td><td>-894.7</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>6.</td><td>ENSP00000379735
| mrm | SNAP | pSYT |</td><td>243</td><td>-854.6</td><td>●</td><td>TP53:p, tumor protein p53</td></tr>
<tr><td>7.</td><td>ENSP00000410739
| mrm | SNAP | pSYT |</td><td>590</td><td>-653.2</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>8.</td><td>ENSP00000425104
| mrm | SNAP | pSYT |</td><td>620</td><td>-568.6</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>9.</td><td>ENSP00000361186
| mrm | SNAP | pSYT |</td><td>1370</td><td>-380.6</td><td>●</td><td>TP53RK:p, TP53 regulating kinase [Source:HGNC Symbol;Acc:16197]</td></tr>
<tr><td>10.</td><td>ENSP00000424104
| mrm | SNAP | pSYT |</td><td>392</td><td>-378.8</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>11.</td><td>ENSP00000423862
| mrm | SNAP | pSYT |</td><td>420</td><td>-333</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>12.</td><td>ENSP00000437792
| SNAP | pSYT |</td><td>80</td><td>-333</td><td>●</td><td>TP53:p, tumor protein p53</td></tr>
<tr><td>13.</td><td>ENSP00000426252
| mrm | SNAP | pSYT |</td><td>211</td><td>-303.2</td><td>●</td><td>TP53:p, tumor protein p53 [Source:HGNC Symbol;Acc:11998]</td></tr>
<tr><td>14.</td><td>ENSP00000394195</td><td>132</td><td>-204.8</td><td>●</td><td>TP53:p, tumor protein p53</td></tr>
</table>

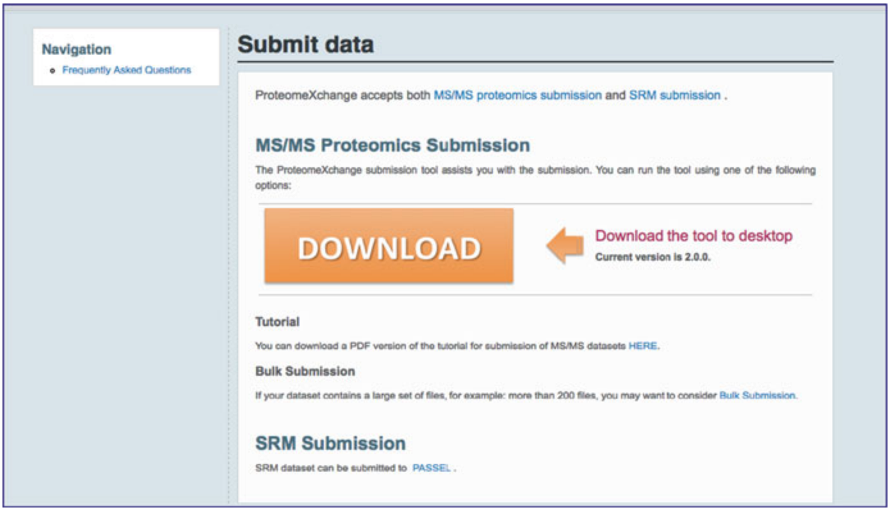
Fig. 4.4 GPM Query: Records related with protein TP53. In contrast to both PRIDE and PeptideAtlas, it is not possible to enquire the database using the UniProtKB ID. In this case the accession is done by the gene ID

MS/MS data can be submitted to the gpmDB via the "simple search page" (http://human.thegpm.org/tandem/thegpm_tandem.html). Once the data has been processed via X!Tandem, users can choose whether or not to publish their data to gpmDB, including with restricted accession (private submission) where only allow user should revise the data.

Furthermore, as PeptideAtlas does, GPM also supports proteomics targeted experiments using SRM.

4.3.2 Proteomics Data Submission to Public Repositories

To illustrate the submission process, here we show the different steps to submit a simple MS/MS dataset to ProteomeXchange (PX). Protein/Peptide identification by mass spectrometry involves the use of tandem mass spectrometry, also known as MS/MS or MS^2 experiments. Below you will find a simple guide to share these kind of data through PX.



Submit data

ProteomeXchange accepts both MS/MS proteomics submission and SRM submission .

MS/MS Proteomics Submission

The ProteomeXchange submission tool assists you with the submission. You can run the tool using one of the following options:

Download the tool to desktop
Current version is 2.0.0.

Tutorial

You can download a PDF version of the tutorial for submission of MS/MS datasets HERE.

Bulk Submission

If your dataset contains a large set of files, for example: more than 200 files, you may want to consider Bulk Submission.

SRM Submission

SRM dataset can be submitted to PASSEL .

Fig. 4.5 ProteomeXchange submission webpage. Here any user could find the indications to submit a dataset to ProteomeXchange. To illustrate the book's example, the PX submission tool has been downloaded just clicking on the banner

1. **Create a PRIDE account**

   If it is the first time you try to submit data, you have to create a PRIDE account. At http://www.ebi.ac.uk/pride/archive/register you will able to do it by just introducing the following parameters: email, name and affiliation.
2. **Download and run PX submission tool**

   The PX submission tool is a desktop application (http://www.proteomexchange.org/submission) that handles MS/MS proteomics data submissions to ProteomeXchange (Fig. 4.5).

Once the application is downloaded, you will be able to upload an experiment through partial or complete submission. Although, the raw data is mandatory for both cases, the difference relays on the results data. Meanwhile for complete submission, a standard XML file is required (PRIDE XML or mzIdentML), partial submission could be done using the search engine output (f.i. search engine's result file). To guide the reader, we will do a complete submission:

(a) To start, you will select complete submission from main interface (Fig. 4.6).
(b) Afterwards, you will login into the system, using your username and password.

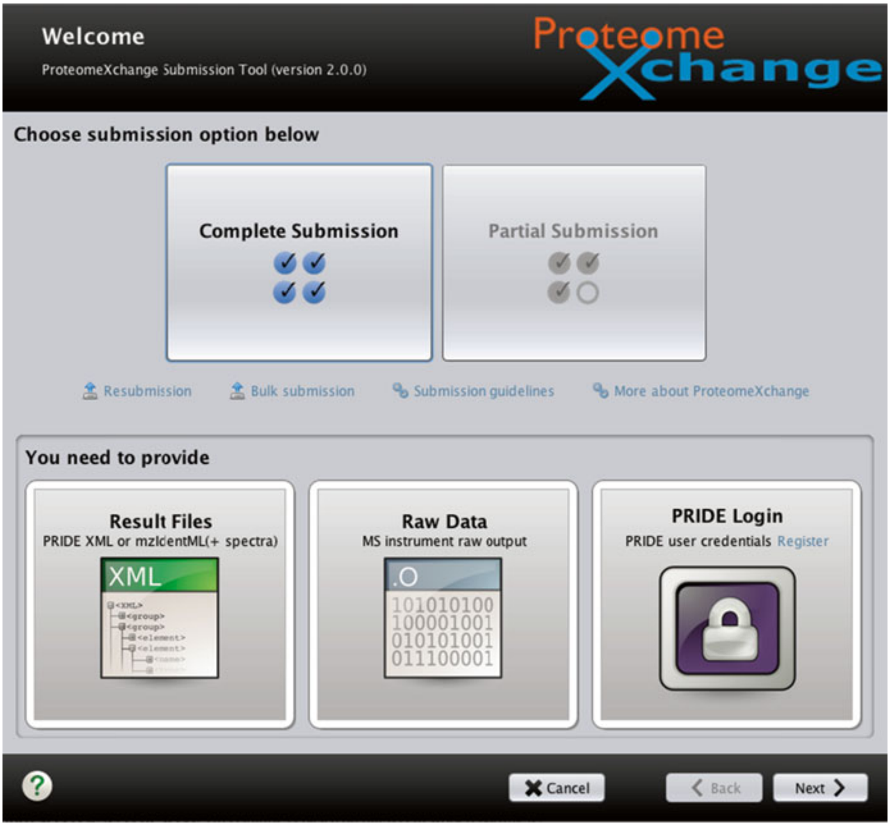
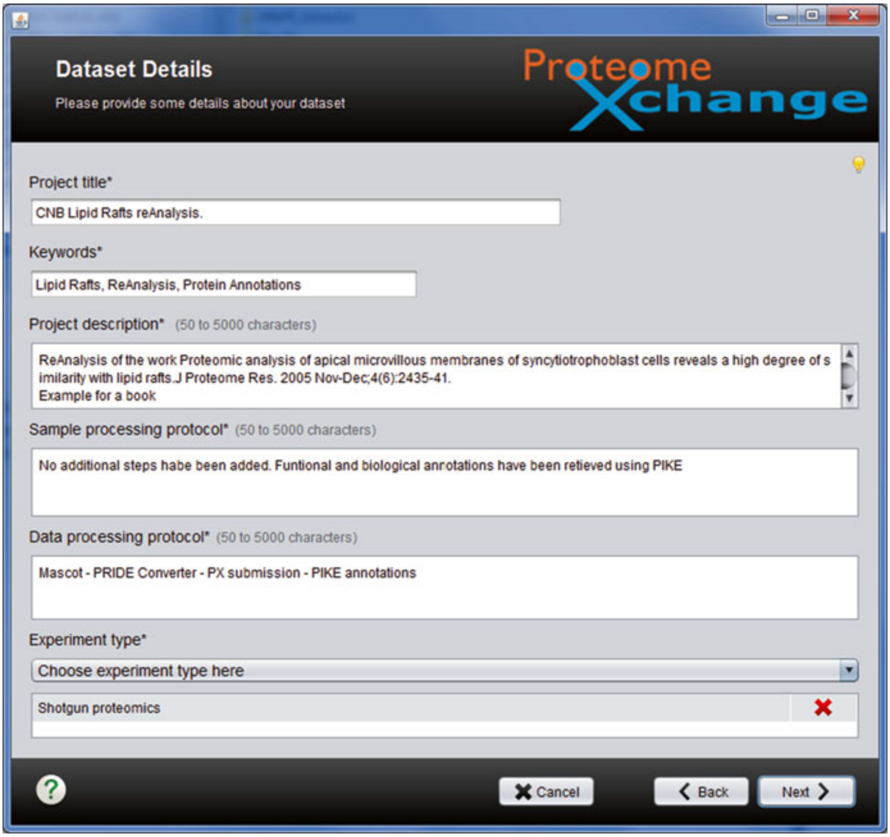


Fig. 4.6 PX submission tool's main interface. First, user should select submission option. For a complete submission is mandatory to enclose a result file formatted as PRIDE XML or mzIdentML plus the unprocessed raw data

(c) Next, a brief description of the experiment is required. You have to take your time to complete it because these metadata will help to other users to retrieve your project from the whole repository. The form asks data for the project title (should be self-descriptive enough), keywords, project description, protocol for both sample processing and data analysis and experiment type (Fig. 4.7).

(d) Next, submission files are required. Several kinds of files can be uploaded, but following with our pipeline, only raw data and results file (PRIDE XML or mzidentml) are mandatory. To illustrate this example (Fig. 4.8), raw data (non-processed peak lists), peak lists (mgf processed peak lists formatted), search data (mascot output .dat file) and result (PRIDE XML) have been added for this



Dataset Details

Please provide some details about your dataset

Project title*

CNB Lipid Rafts reAnalysis.

Keywords*

Lipid Rafts, ReAnalysis, Protein Annotations

Project description* (50 to 5000 characters)

ReAnalysis of the work Proteomic analysis of apical microvillous membranes of syncytiotrophoblast cells reveals a high degree of s
imilarity with lipid rafts.J Proteome Res. 2005 Nov-Dec;4(6):2435-41.
Example for a book

Sample processing protocol* (50 to 5000 characters)

No additional steps habe been added. Funtional and biological anrotations have been retieved using PIKE

Data processing protocol* (50 to 5000 characters)

Mascot - PRIDE Converter - PX submission - PIKE annotations

Experiment type*

Choose experiment type here

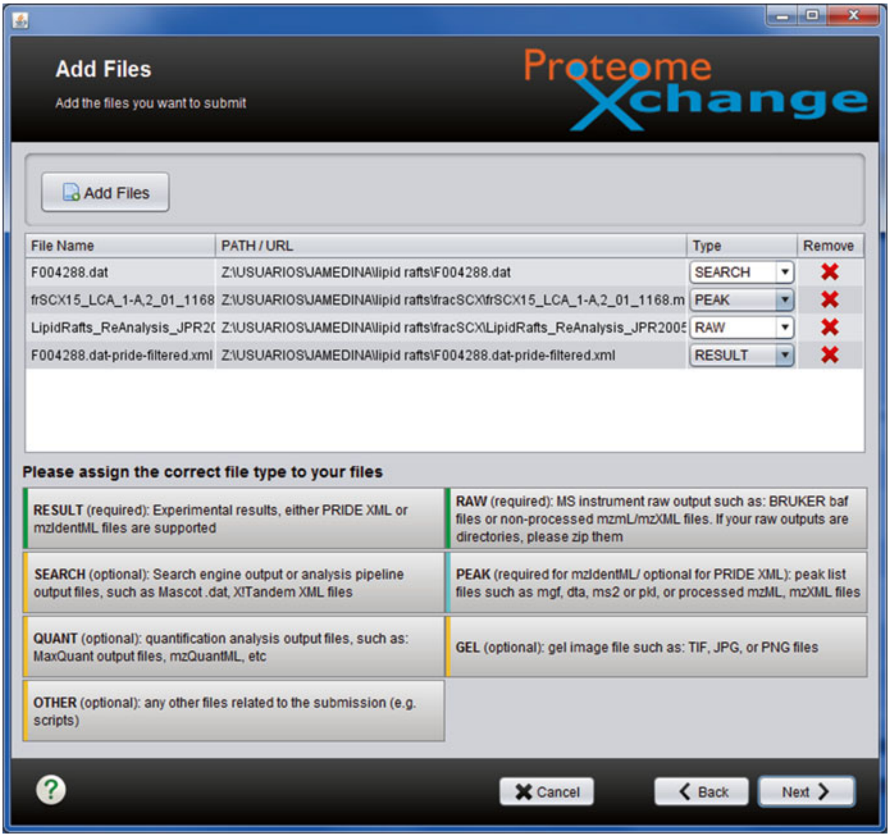
Shotgun proteomics

Fig. 4.7 First, experiment metadata is required. The user enters descriptions about how the experiment was performed, including a brief summary regarding the analytical or the data processing protocols. It provides an understandable explanation of the experiment

submission. The PX application will check the files in the next step, before the submission.

(e) Finally, the metadata regarding the experiment sample and protocol could be added, before to finish the submission (Fig. 4.9). Once the submission is finished, an email regarding the submission will be sent.

This example has been recorded with the ProteomeXchange ID PXD000744 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD000744) and DOI 10.6019/PXD000744 for further revisions by readers.



Add Files

Add the files you want to submit

Add Files

File Name	PATH / URL	Type	Remove
F004288.dat	Z:\USUARIOS\JAMEDINA\lipid rafts\F004288.dat	SEARCH	✖
frSCX15_LCA_1-A,2_01_1168	Z:\USUARIOS\JAMEDINA\lipid rafts\fracSCX\frSCX15_LCA_1-A,2_01_1168.m	PEAK	✖
LipidRafts_ReAnalysis_JPR2(	Z:\USUARIOS\JAMEDINA\lipid rafts\fracSCX\LipidRafts_ReAnalysis_JPR2005	RAW	✖
F004288.dat-pride-filtered.xml	Z:\USUARIOS\JAMEDINA\lipid rafts\F004288.dat-pride-filtered.xml	RESULT	✖

Please assign the correct file type to your files

RESULT (required): Experimental results, either PRIDE XML or mzIdentML files are supported

RAW (required): MS instrument raw output such as: BRUKER baf files or non-processed mzmL/mzXML files. If your raw outputs are directories, please zip them

SEARCH (optional): Search engine output or analysis pipeline output files, such as Mascot .dat, X!Tandem XML files

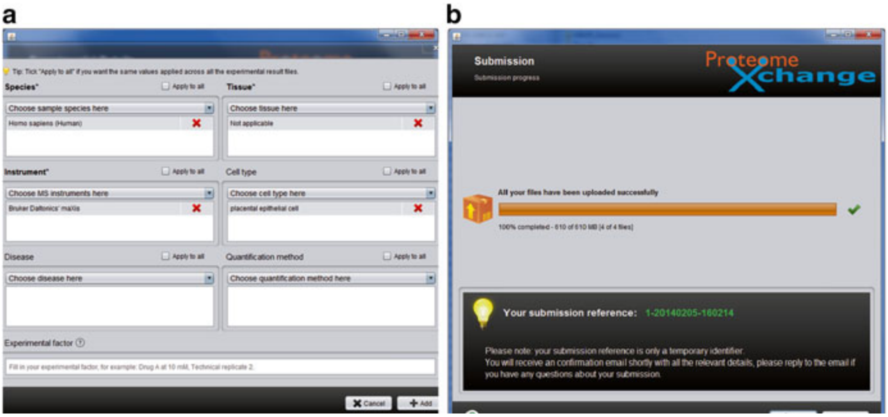
PEAK (required for mzIdentML/ optional for PRIDE XML): peak list files such as mgf, dta, ms2 or pkl, or processed mzML, mzXML files

QUANT (optional): quantification analysis output files, such as: MaxQuant output files, mzQuantML, etc

GEL (optional): gel image file such as: TIF, JPG, or PNG files

OTHER (optional): any other files related to the submission (e.g. scripts)

Fig. 4.8 Next, the files are entered. In this example, the mandatory files (results and raw) are attached together the mascot results (.dat file) and the processed peak list (.mgf file) for those that only wants to repeat the peptide and protein identification step. Furthermore, a list containing the identified protein has also submitted



Tip: Tick "Apply to all" if you want the same values applied across all the experimental result files.

Species* □ Apply to all

Choose sample species here

Homo sapiens (Human) ✖

Tissue* □ Apply to all

Choose tissue here

Not applicable ✖

Instrument* □ Apply to all

Choose MS instruments here

Bruker Daltonics' maXis ✖

Cell type □ Apply to all

Choose cell type here

placental epithelial cell ✖

Disease □ Apply to all

Choose disease here

Quantification method □ Apply to all

Choose quantification method here

Experimental factor

Fill in your experimental factor, for example: Drug A at 10 mM, Technical replicate 2

Submission

Submission progress

All your files have been uploaded successfully

100% completed - 610 of 610 MB [4 of 4 files]

Your submission reference: 1-20140205-160214

Please note: your submission reference is only a temporary identifier.
You will receive an confirmation email shortly with all the relevant details, please reply to the email if you have any questions about your submission.

Fig. 4.9 Submission's final steps: (**a**) Details regarding the sample and the mass spectrometers. (**b**) Submission status

Table 4.1 Some of the most valuable annotation databases for proteomics experiments

Annotation type	URL
Additional features and keywords	http://www.uniprot.org/
Taxonomy	http://www.ncbi.nlm.nih.gov/Taxonomy/
Gene Ontology	
Amigo	http://amigo.geneontology.org/
Metabolic pathways	
KEGG	http://www.genome.jp/kegg/pathway.html
Diseases	
OMIM	http://omim.org/
Protein interactions	
STRING	http://string-db.org/
IntAct	http://www.ebi.ac.uk/intact/main.xhtml
PTMs	
Phosida	http://www.phosida.com/
Phosphosite	http://www.phosphosite.org/
Protein families, domains and functional sites	
Interpro	http://www.ebi.ac.uk/interpro/
Pharma drugs	
PharmaGKB	http://www.pharmgkb.org/

4.3.3 Proteomics Related Data: Human Resources

Analytical results show the evidence of the detected proteins regarding to a set of samples. However, to get further knowledge is important to know their biological context. To improve the knowledge of a concrete protein, information regarding to the metabolic pathways (KEGG, Biocarta) in which these proteins can be associated, their potential and/or reported molecular interactions (IntAct, STRING), possible role on diseases (OMIM), potential posttranslational modifications (Phosphosite, Phosida) or biological functions or molecular process (Gene Ontology, GO) can be compiled through curated and freely available repositories (see Table 4.1).

Here we describe a subset of the most accessed databases to retrieve some of the key features that may guide to understand the biological role played by a set of identified proteins. In addition to the protein related information, all of the following significant repositories offer a close connection with the Ensembl project (http://www.ensembl.org/) for crossing the available data with the genome databases. It provides a framework where genomics and proteomics background could be linked and matched.

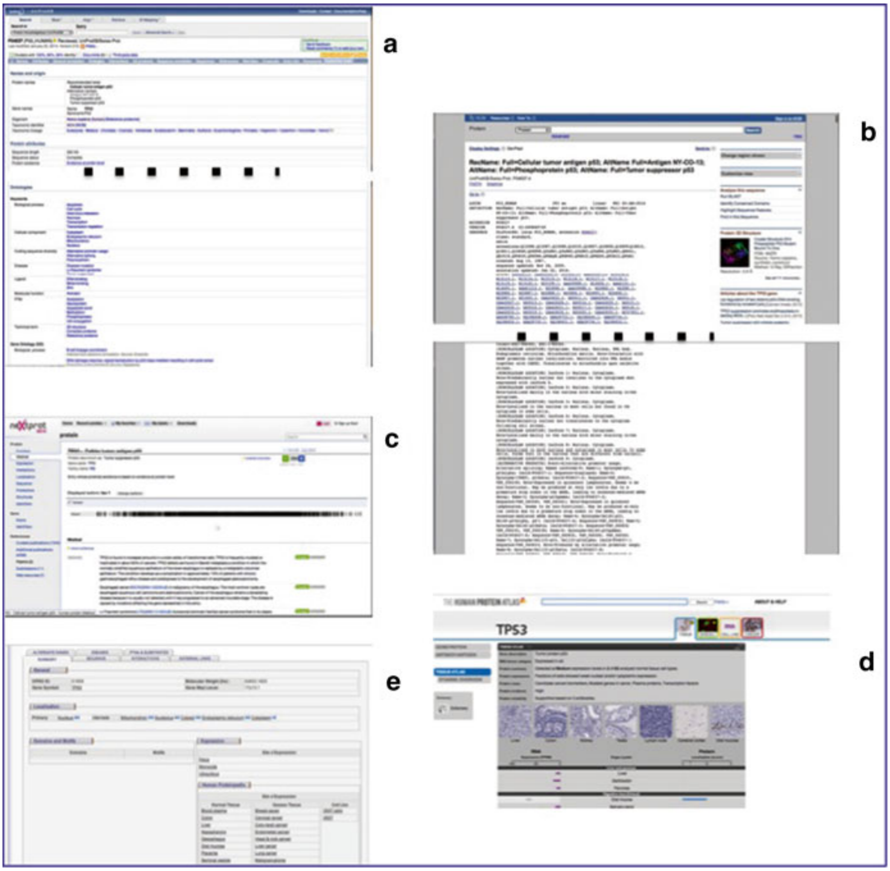


Fig. 4.10 Different views offered by UniProtKB (**a**), NCBI (**b**), neXtProt (**c**), Human Protein Atlas (**d**) and Human Protein Reference Databases (**e**) for P53 protein. Data regarding GO ontology, keywords, diseases, interactions, cell types, tissues, etc… are offered by these databases to complement the analytical results

4.3.3.1 Universal Protein Resource Knowledge Base, UniProtKB (http://www.uniprot.org/)

The UniProt Knowledgebase (UniProtKB) as a central repository collects functional information with a very well curated, consistent and rich annotation of proteins. The enclosed information is shown according to two categories: First, the mandatory data for each UniProtKB entry: the amino acid sequence, protein name or description, taxonomic data and citation information. These data are useful to define without any doubt different entries. Second, additional annotations, as much as possible, are added. This includes widely accepted biological ontologies, such as Gene Ontology, cross-references to other repositories, and faultless indications of the quality of annotation in the form of evidence attribution of experimental and computational data (Fig. 4.10a).

The UniProtKB consists of two sections: a section containing manually-annotated and curated records with information extracted from literature and evaluated through computational analysis, and a section with only computationally evaluated records that are awaiting for fully manual annotation or experimental evidences. The two sections are referred to as “UniProtKB/Swiss-Prot” (reviewed, manually annotated) and “UniProtKB/TrEMBL” (unreviewed, automatically annotated) respectively. Currently is maintained by the EBI, the Swiss Bioinformatics Institute (SIB) and the Protein Information Resource (PIR).

By January 2014, UniProtKB/Swiss-Prot contains 542,258 protein entries, derived from the sum of 225,339 references and linked to 13,052 species. The human case is the most representative taxonomy with 20,272 entries. To enquire UniProtKB, user could go to the website and perform a simple query with only the protein accession code or protein name. All the available information is depicted as a table.

4.3.3.2 National Center for Biotechnology Information, NCBI (http://www.ncbi.nlm.nih.gov)

The NCBI offers a wide portfolio of resources related with molecular biology information for advancing science and health by providing access to biomedical and genomic information. NCBI’s mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules.

In terms of databases and tools, NCBI joins around 70 databases and more than 50 tools including some of outstanding focalized repositories such as the Genome Reference Consortium, UniGene, PubMed Central, the most accessed library for disease’s data (OMIM), several BLAST distributions or different utilities to browse genomes are some of the most well-known offered aids (Fig. 4.10b).

4.3.3.3 A Knowledge Platform for Human Proteins: neXtprot (http://www.nextprot.org/)

Developed in collaboration between the Swiss Institute of Bioinformatics (SIB) and Geneva Bioinformatics (GeneBio) SA, neXtprot is a comprehensive resource focussed on human-centric discovery through protein-related data. Based on UniProtKB/Swiss-Prot human proteins (nextprot sums the same human protein entries), it increases the available data linking the UniProtKB annotations to some of the most accessed public repositories related to omics research. A complete list of the data resources integrated within neXtprot is available at: http://www.nextprot.org/db/statistics/release.

Through the integrated resources, we would like to highlight the integration with COSMIC and PeptideAtlas. The first, COSMIC is an important database for cancer research. It stores and displays somatic mutation information and related details relating to human cancers. PeptideAtlas, was described previously as part of ProteomeXchange Consortium, being one of the most important repositories for MS-based proteomics data (specially SRM/MRM datasets). Furthermore, It is also connected with Human Protein Atlas, below described.

Perform queries in neXtprot is quite similar to UniProt. Using the protein name or protein accession code (UniProtKB IDs are allowed), user can navigate through the different sections to check the available information (Fig. 4.10c).

4.3.3.4 The Human Protein Atlas (http://www.proteinatlas.org/)

The Human Protein Atlas brings a systematic exploration of the human proteome using Antibody-Based Proteomics. This is accomplished by combining high-throughput generation of affinity-purified antibodies with protein profiling in a multitude of tissues and cells assembled in tissue microarrays. In detail, millions of high-resolution images showing the spatial distribution of proteins in 44 different normal human tissues and 20 different cancer types, as well as 46 different human cell lines are shown. As the previous systems, the database can be interrogated using UniProtKB IDs.

The database has been developed in a gene-centric manner with the inclusion of all human genes predicted from genome initiatives. Search functionalities allow for complex queries regarding protein expression profiles, protein classes and chromosome location. To provide a high confidence data, each antibody is enclosed with its correspondence application-specific validation, including immunohistochemistry, western blot analysis and, for a large fraction, a protein array assay and immunofluorescent-based confocal microscopy (Fig. 4.10d).

4.3.3.5 The Human Protein Reference Database HPRD (http://www.hprd.org/)

The Human Protein Reference Database (HPRD) represents a centralized platform depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome (Fig. 4.10e). All the information in HPRD has been manually curated from the literature by experts. By January 2014, 30,047 protein entries are described (including isoforms), including more than 41,000 protein-protein interactions and around 94,000 PTMs. In addition, HPRD integrates data deposited in Human Proteinpedia along with the existing literature curated information in the context of an individual protein. All the public data contributed to Human Proteinpedia can be queried, viewed and downloaded at http://www.humanproteinpedia.org/. UniProtKB IDs may be used to get the information regarding a concrete entry.

4.4 Tools for Automatic Data Retrieving

In most of the cases, proteome-wide experiments generate a complex array of information represented as a set of identified, characterized or differentially expressed proteins. Hundreds or thousands of proteins can be reported within a single experiment. Linking these proteins to the available information stored in some of the most popular databases, such as the previously described, can offer valuable information for proteome researchers. This may provides wealth information about the biological role of the proteins that eventually could be correlated with the specific topic of the experiment.

However, retrieving the content, making annotations and establishing relationships among the entries through various databases means to deal with different uses of nomenclature, data formats, and accession modes. Furthermore, this has been a hurdle to garner functional and non-redundant information. Due to the complexity of this task and the huge amount of data available, it is not feasible to gather this information by hand, making it necessary to have automatic aids. Here, we describe some of the most popular tools in order to retrieve the available information related to a set of genes and proteins.

4.4.1 Genomics and Transcriptomic Data: Babelomics, GeneCodis

Babelomics (http://babelomics.bioinfo.cipf.es/) (Medina et al. 2010) is an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Version 4.3.0 (2013) of Babelomics integrates primary (normalization, calls, etc.) and secondary (signatures, predictors, associations, clustering, etc.) analysis tools within an environment that mainly allows relating genomic data and/or interpreting them by means of different functional enrichment or gene set methods. Such interpretation is made using functional definitions such as GO, KEGG, Biocarta, and also regulatory information from Transfac and other levels of regulation such as miRNA-mediated interference, protein-protein interactions, text-mining module definitions and the possibility of producing de novo annotations through the Blast2GO system.

Complementary to Babelomics, **GeneCodis3** (http://genecodis.cnb.csic.es/) (Tabas-Madrid et al. 2012) is a web-based tool for the ontological analysis of large lists of genes. It can be used to determine biological annotations or combinations of annotations that are significantly associated to a list of genes under study with respect to a reference list. As well as single annotations, this tool allows users to simultaneously evaluate annotations from different sources, for example Biological Process and Cellular Component categories of Gene Ontology.

In addition, all the analysis could be accessed through a set of web services in favour of third-party software development.

4.4.2 Genomics and Proteomics Enrichment: DAVID

The Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) (da Huang et al. 2009) provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes or proteins. Although it is mainly focused on gene analysis, for any given gene or protein list, DAVID tools are able to: Identify enriched biological themes, particularly GO terms, discover enriched functional-related gene groups, cluster redundant annotation terms, visualize genes on BioCarta or KEGG pathway maps, list interacting proteins, link gene-disease associations or highlight protein functional domains and motifs.

Despite of these analyses are available for genes and proteins, proteomics workflow is barely supported by this tool and only using UniProtKB accession codes. It forces to users to previously translate their protein accession codes to UniProtKB IDs when other databases were used during identification phase. Furthermore, enrichment process is done comparing the enquired proteins against the information derived from reference genomes. It provides a partial interpretation of the over-represented proteins due to there are no evidences regarding the detection of some proteins from real expressed genes.

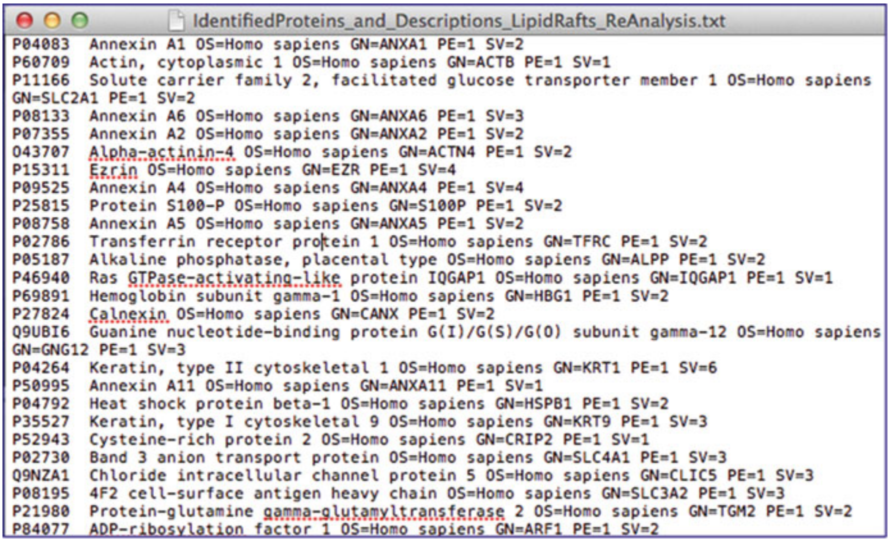
By other hand, DAVID analysis could be integrated within third-party tools thorough its web Service (Jiao et al. 2012). It has been largely tested for the functional interpretation of large lists of genes/proteins.

4.4.3 Proteomics Data Integration: PIKE

Although previous tools show an outstanding performance for the genomics scenario, they do not deal with the most frequent proteomics workflows. This is a critical issue when a proteomic environment is evaluated due to the assortment of databases currently in use for protein identification. The restriction to a very limited number of sequence database codes greatly limits its usefulness and applicability. Protein Information and Knowledge Extractor (PIKE, http://proteo.cnb.csic.es/pike) (Medina-Aunon et al. 2010) admits all available protein identifiers used in the most common databases: Uniprot, NCBI nr, GenBank, EMBL, PDB, DDBJ and Entrez Gene ID. Using any of these identifiers as input, PIKE checks the information stored in a set of protein databases and systematically extracts and reports in real-time, non-redundant functional and biological information. Finally, relevant information is reported to the user in a wide range of standard output formats that can be viewed, exported or downloaded.

4.5 Example Using PIKE

Here, we describe how to use PIKE. Through a few single steps, you will find how to complement the analytical results (encoded by a list of protein identifiers) with the most recent available biological and functional annotations.



IdentifiedProteins_and_Descriptions_LipidRafts_ReAnalysis.txt

```
P04083  Annexin A1 OS=Homo sapiens GN=ANXA1 PE=1 SV=2
P60709  Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1
P11166  Solute carrier family 2, facilitated glucose transporter member 1 OS=Homo sapiens
GN=SLC2A1 PE=1 SV=2
P08133  Annexin A6 OS=Homo sapiens GN=ANXA6 PE=1 SV=3
P07355  Annexin A2 OS=Homo sapiens GN=ANXA2 PE=1 SV=2
O43707  Alpha-actinin-4 OS=Homo sapiens GN=ACTN4 PE=1 SV=2
P15311  Ezrin OS=Homo sapiens GN=EZR PE=1 SV=4
P09525  Annexin A4 OS=Homo sapiens GN=ANXA4 PE=1 SV=4
P25815  Protein S100-P OS=Homo sapiens GN=S100P PE=1 SV=2
P08758  Annexin A5 OS=Homo sapiens GN=ANXA5 PE=1 SV=2
P02786  Transferrin receptor protein 1 OS=Homo sapiens GN=TFRC PE=1 SV=2
P05187  Alkaline phosphatase, placental type OS=Homo sapiens GN=ALPP PE=1 SV=2
P46940  Ras GTPase-activating-like protein IQGAP1 OS=Homo sapiens GN=IQGAP1 PE=1 SV=1
P69891  Hemoglobin subunit gamma-1 OS=Homo sapiens GN=HBG1 PE=1 SV=2
P27824  Calnexin OS=Homo sapiens GN=CANX PE=1 SV=2
Q9UBI6  Guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-12 OS=Homo sapiens
GN=GNG12 PE=1 SV=3
P04264  Keratin, type II cytoskeletal 1 OS=Homo sapiens GN=KRT1 PE=1 SV=6
P50995  Annexin A11 OS=Homo sapiens GN=ANXA11 PE=1 SV=1
P04792  Heat shock protein beta-1 OS=Homo sapiens GN=HSPB1 PE=1 SV=2
P35527  Keratin, type I cytoskeletal 9 OS=Homo sapiens GN=KRT9 PE=1 SV=3
P52943  Cysteine-rich protein 2 OS=Homo sapiens GN=CRIP2 PE=1 SV=1
P02730  Band 3 anion transport protein OS=Homo sapiens GN=SLC4A1 PE=1 SV=3
Q9NZA1  Chloride intracellular channel protein 5 OS=Homo sapiens GN=CLIC5 PE=1 SV=3
P08195  4F2 cell-surface antigen heavy chain OS=Homo sapiens GN=SLC3A2 PE=1 SV=3
P21980  Protein-glutamine gamma-glutamyltransferase 2 OS=Homo sapiens GN=TGM2 PE=1 SV=2
P84077  ADP-ribosylation factor 1 OS=Homo sapiens GN=ARF1 PE=1 SV=2
```

Fig. 4.11 PIKE's input. List containing the 125 identified proteins derived by the reported experiment. Just including the protein accession codes it is possible to run PIKE

4.5.1 How to Start?

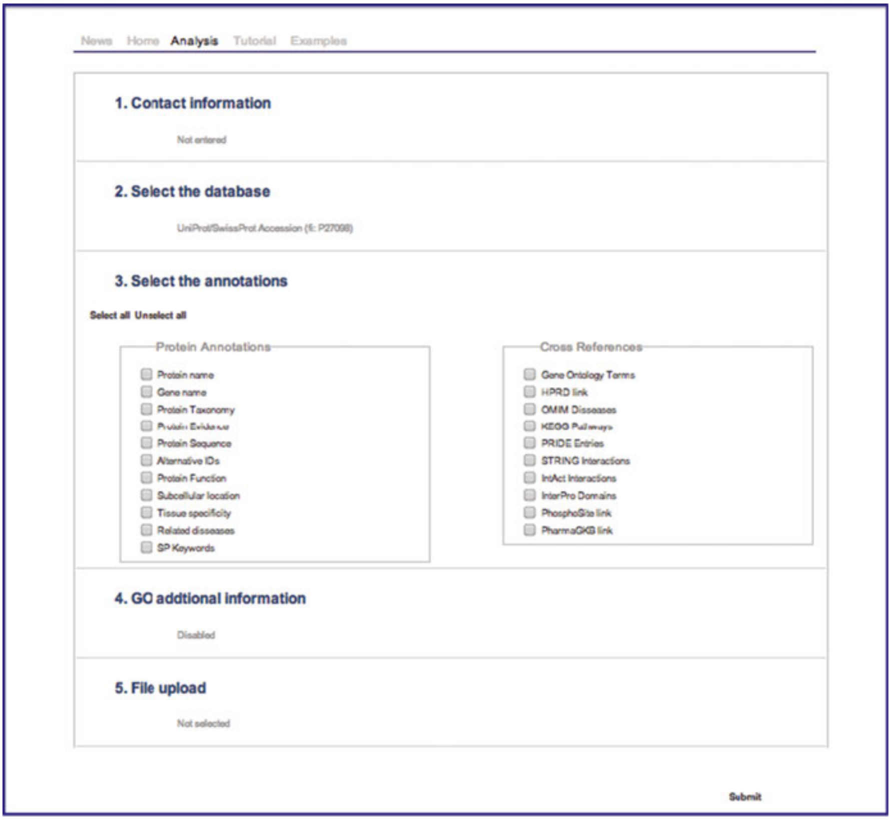
For this example, all the identified proteins derived from the experiment described at 3.2 have been used as input. The protein accession codes (UniProtKB IDs) were collected in a single text file. The input arranges a row for each protein accession code and optimally its protein name (Fig. 4.11). A total of 125 proteins are included in the input file.

Although UniProtKB IDs are closely the most frequent codes reported within results, to support the most used databases in proteomics, PIKE admits other proteins accession codes. To ensure the accuracy of the cross-reference protein list, PIKE's algorithm introduces an additional step during the execution and checks on the fly the available data at Protein Information Resource website (http://pir.georgetown.edu/) to confirm the cross-references.

4.5.2 Entering Your Query

Throughout the main interface (Fig. 4.12) the user has to enter a few and simple parameters to run the query.

1. User name and e-mail address. The reason for requiring this information is to allow the results of a search to be returned by email. (Optional)



1. Contact information

Not entered

2. Select the database

UniProt/SwissProt Accession (fi: P27098)

3. Select the annotations

Select all Unselect all

Protein Annotations

□ Protein name
□ Gene name
□ Protein Taxonomy
□ Protein Evidence
□ Protein Sequence
□ Alternative IDs
□ Protein Function
□ Subcellular location
□ Tissue specificity
□ Related diseases
□ SP Keywords

Cross References

□ Gene Ontology Terms
□ HPRD link
□ OMIM Diseases
□ KEGG Pathways
□ PRIDE Entries
□ STRING Interactions
□ IntAct Interactions
□ InterPro Domains
□ PhosphoSite link
□ PharmaGKB link

4. GO addtional information

Disabled

5. File upload

Not selected

Fig. 4.12 PIKE's main interface. User name and/or e-mail, database used to obtain the protein list, annotations and finally the input file are enough to run a query

2. Specify the Database used to obtain the protein list: Uniprot, NCBI nr, GenBank, EMBL, PDB, DDBJ, Entrez Gene ID.
3. Select one, a subset or all of the functional fields of interest, according to the objectives of the experiment. Most of them are compiled in Table 4.1.
4. Select the specific level of representation for Gene Ontology clustering (Optional).
5. Upload the input protein list. Text or PRIDE XML files

The fourth parameter is designed for small queries where the user is interested in getting the whole information regarding the GO hierarchy related to the input protein list.

To perform the query, the user only has to click on submit to start the extraction of the data. To sum up, this process firstly, validate the input file by checking the protein accession codes. Afterwards, PIKE begins its execution across the three modules that the main algorithm has been divided. In that way, the first module (Workflow Manager Module, WFMM) checks the input protein list and selects feasible sources (functional databases) to obtain the required information. Afterward, the second module (Information Retrieving Module, IRM) follows the previously defined routes and retrieves the desired information. And then, the third module (File Manager Module, FMM) compiles all information and generates the result document and output files.

4.5.3 Showing the Results

Once the analysis process has been initiated, based on the size of the protein list, PIKE selects the execution mode. This process will allow the user to monitor the retrieved information on the fly – real-time mode – or to receive the information by e-mail – default mode. Regardless of the execution mode used, an output file (as XHTML web page) is reported (Fig. 4.13) displaying the following elements:

The XHTML or screen display where the proteins are depicted within in a table. Each protein heads a row and each field of interest is a column. And the files where the results containing the required information are compressed and downloaded in a .zip file. The results are exported according to the formats defined in the FMM module: PRIDE XML, Comma Separated Values (CSV) or plain text, including clusters of the following annotations: Swiss-Prot Keywords, OMIN terms, KEGG pathways and Gene Ontology classification.

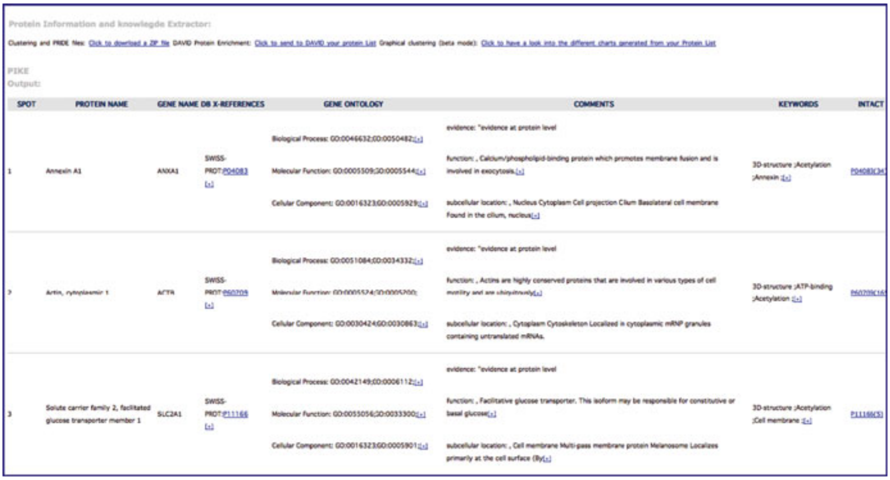
4.5.4 Poring Over the Data

One of the most recent features included into PIKE is to show the different clusters. Using Google's aids, the different graphical distributions according the main areas of interest are depicted. Figure 4.14 shows some examples of these views.

Furthermore, PIKE enables a direct link to DAVID. Using its web service it is possible to link the enquired proteins to DAVID enrichment service to provide a statistical measure of the protein over-representation considering the human genome as background (Fig. 4.15).

4.5.5 Integrating the Results

Regarding export formats, one of the advantages of PIKE is the possibility to integrate with existing laboratory workflows implementing the standard format PRIDE XML. It supports those users who want to supplement the analytical results with the available annotations. Using this format, It is therefore easy to extend existing laboratory workflows by connecting the experimental information stored in PRIDE XML files with the information available in public databases gathered by PIKE, within a single file.



Protein Information and knowlegde Extractor:

Clustering and PRIDE files: Click to download a ZIP file DAVID Protein Enrichment: Click to send to DAVID your protein List Graphical clustering (beta mode): Click to have a look into the different charts generated from your Protein List

PIKE
Output:

SPOT	PROTEIN NAME	GENE NAME	DB X-REFERENCES	GENE ONTOLOGY	COMMENTS	KEYWORDS	INTACT
1	Annexin A1	ANXA1	SWISS-PROT:P04083 [+]	Biological Process: GO:0046632;GO:0050482;[+] Molecular Function: GO:0005509;GO:0005544[+] Cellular Component: GO:0016323;GO:0005929;[+]	evidence: "evidence at protein level function: , Calcium/phospholipid-binding protein which promotes membrane fusion and is involved in exocytosis.[+] subcellular location: , Nucleus Cytoplasm Cell projection Cilium Basolateral cell membrane Found in the cilium, nucleus[+]	3D-structure ;Acetylation ;Annexin ;[+]	P04083(34
2	Actin, cytoplasmic 1	ACTB	SWISS-PROT:P60709 [+]	Biological Process: GO:0051084;GO:0034332;[+] Molecular Function: GO:0005524;GO:0005200; Cellular Component: GO:0030424;GO:0030863;[+]	evidence: "evidence at protein level function: , Actins are highly conserved proteins that are involved in various types of cell motility and are ubiquitously[+] subcellular location: , Cytoplasm Cytoskeleton Localized in cytoplasmic mRNP granules containing untranslated mRNAs.	3D-structure ;ATP-binding ;Acetylation ;[+]	P60709(16
3	Solute carrier family 2, facilitated glucose transporter member 1	SLC2A1	SWISS-PROT:P11166 [+]	Biological Process: GO:0042149;GO:0006112;[+] Molecular Function: GO:0055056;GO:0033300;[+] Cellular Component: GO:0016323;GO:0005901;[+]	evidence: "evidence at protein level function: , Facilitative glucose transporter. This isoform may be responsible for constitutive or basal glucose[+] subcellular location: , Cell membrane Multi-pass membrane protein Melanosome Localizes primarily at the cell surface (By[+]	3D-structure ;Acetylation ;Cell membrane ;[+]	P11166(5)

Fig. 4.13 PIKE's xhtml output. All the available annotations are depicted in a single view. Each row displays the data regarding a single protein. In addition, there are some links to download the gathered information formatted as PRIDE XML or CSV files, to run DAVID with the input protein list and some graphical views

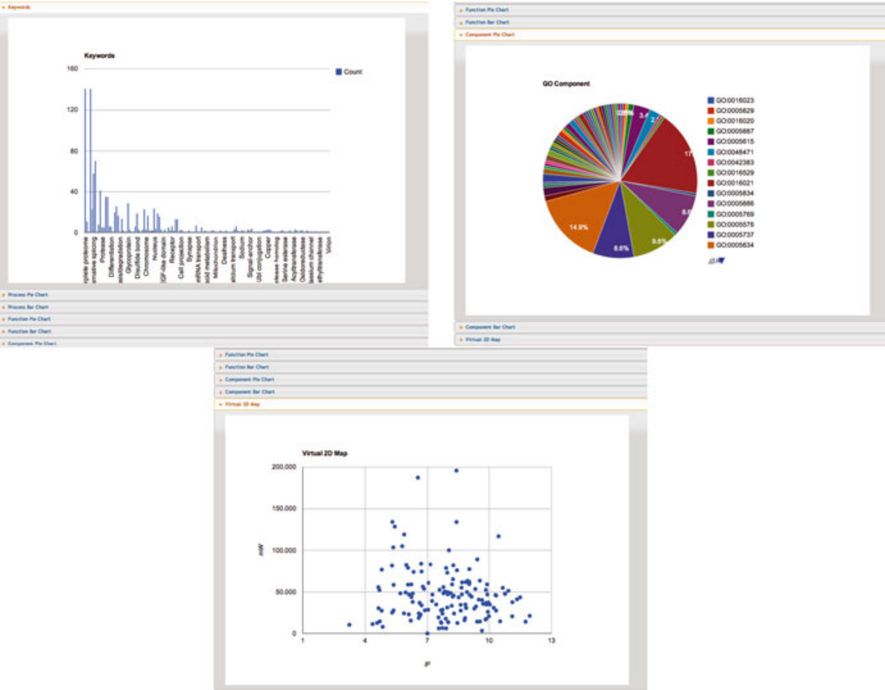
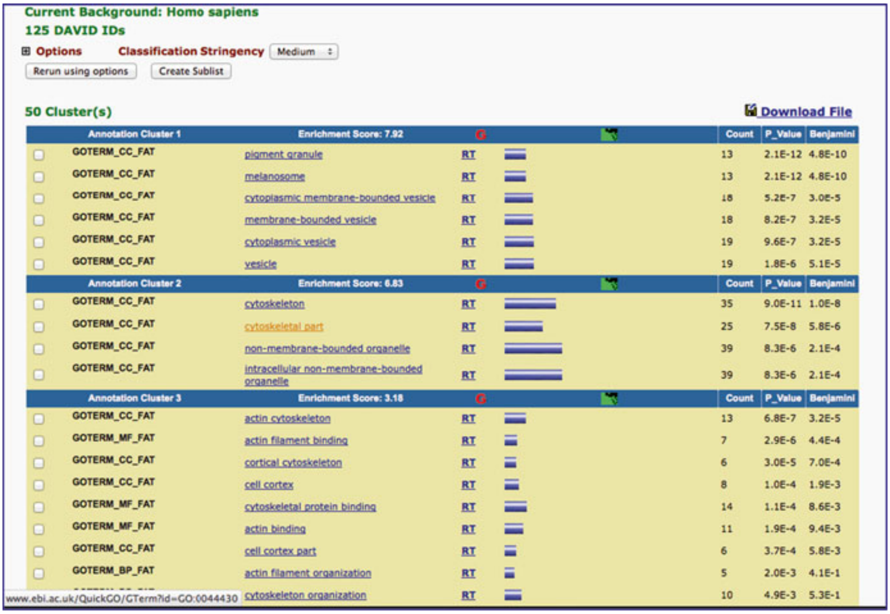


Fig. 4.14 PIKE's graphical views. Distributions regarding Uniprot keywords (**a**), GO ontologies (**b**) or a virtual interpretation of Mw/pI like a 2D Gel are provided to get a more friendly results' view



Current Background: Homo sapiens
125 DAVID IDs

Classification Stringency: Medium

50 Cluster(s)

	Annotation Cluster 1	Enrichment Score: 7.92	G	Count	P_Value	Benjamini
□	GOTERM_CC_FAT	pigment granule	RT	13	2.1E-12	4.8E-10
□	GOTERM_CC_FAT	melanosome	RT	13	2.1E-12	4.8E-10
□	GOTERM_CC_FAT	cytoplasmic membrane-bounded vesicle	RT	18	5.2E-7	3.0E-5
□	GOTERM_CC_FAT	membrane-bounded vesicle	RT	18	8.2E-7	3.2E-5
□	GOTERM_CC_FAT	cytoplasmic vesicle	RT	19	9.6E-7	3.2E-5
□	GOTERM_CC_FAT	vesicle	RT	19	1.8E-6	5.1E-5
	Annotation Cluster 2	**Enrichment Score: 6.83**	**G**	**Count**	**P_Value**	**Benjamini**
□	GOTERM_CC_FAT	cytoskeleton	RT	35	9.0E-11	1.0E-8
□	GOTERM_CC_FAT	cytoskeletal part	RT	25	7.5E-8	5.8E-6
□	GOTERM_CC_FAT	non-membrane-bounded organelle	RT	39	8.3E-6	2.1E-4
□	GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT	39	8.3E-6	2.1E-4
	Annotation Cluster 3	**Enrichment Score: 3.18**	**G**	**Count**	**P_Value**	**Benjamini**
□	GOTERM_CC_FAT	actin cytoskeleton	RT	13	6.8E-7	3.2E-5
□	GOTERM_MF_FAT	actin filament binding	RT	7	2.9E-6	4.4E-4
□	GOTERM_CC_FAT	cortical cytoskeleton	RT	6	3.0E-5	7.0E-4
□	GOTERM_CC_FAT	cell cortex	RT	8	1.0E-4	1.9E-3
□	GOTERM_MF_FAT	cytoskeletal protein binding	RT	14	1.1E-4	8.6E-3
□	GOTERM_MF_FAT	actin binding	RT	11	1.9E-4	9.4E-3
□	GOTERM_CC_FAT	cell cortex part	RT	6	3.7E-4	5.8E-3
□	GOTERM_BP_FAT	actin filament organization	RT	5	2.0E-3	4.1E-1
	www.ebi.ac.uk/QuickGO/GTerm?id=GO:0044430	cytoskeleton organization	RT	10	4.9E-3	5.3E-1

Fig. 4.15 David analysis. From a direct link included in the results, it is possible to link to DAVID to get a measure of annotations' association

Moreover, the inclusion of controlled vocabularies from PRIDE XML ensures compatibility with other tools and databases. In that way, analogous to PRIDE schema, PIKE divides the retrieved information in two groups: the first may be validated using specific controlled vocabularies (CV) or Ontologies such as Gene Ontology, IntAct, KEGG or OMIN terms while the second is free information without any control of the terms as in the case of UniProtKB/Swiss-Prot manual-annotated comments. This feature allows to semantically validate PIKE's output and to assure the information accuracy. Once PIKE has included the biological information, the generated EBI PRIDE XML file can be submitted to the PRIDE central repository.

4.6 Conclusions

Here, we have highlighted the importance of reporting proteomics experimental data with sufficient detail to provide a pipeline for the replication and reanalysis of proteomics results. This is governed by a cohort of minimal information requirement guidelines (MIAPE), standardized data formats (XML) and semantic validators, all of which proposed by the proteomics community through the HUPO-PSI working groups.

Primarily linked to this MS-based data generation issue, there are proteomics databases (PRIDE, PeptideAtlas, GPMDB), now under the ProteomeXchange Consortium, where either raw data and/or experimental metadata may be collected according to specific deposition rules. These repositories have to be in close relation with other protein genome wide-related databases to get further knowledge of the biological context of a set of identified proteins. Web portals like the UniProt Knowledgebase (UniProtKB), neXtprot, NCBI, as central repositories, collect functional information with a very well curated, consistent and rich annotation of proteins. In addition, information regarding the metabolic pathways (KEGG, Biocarta) in which a set of proteins can be associated, their potential and/or reported molecular interactions (IntAct, STRING), the possible link with diseases (OMIM), the potential posttranslational modifications (Phosphosite, Phosida) or the biological functions or molecular process (Gene Ontology, GO) can all be compiled through curated and freely available repositories. HPA and HPRD may also complement and enrich the functional annotation of proteins.

Finally, freely available web tools, like PIKE (Protein Information and Knowledge Extractor) admits all available protein identifiers from the most common databases (Uniprot, NCBI nr, GenBank, EMBL, PDB, DDBJ and Entrez Gene DB) to systematically extract and report in real-time, non-redundant functional and biological information. This information may be reported to the user in a wide range of standard output formats that can be viewed, exported or downloaded making easily achievable the integration of crude proteomics data with the existing functional annotation of proteins.

Incorporating all these tools and applying the adequate guidelines will be critical to link the current global knowledge generated over these last decades and to provide key information for future experimental queries.

References

Binz PA, Barkovich R, Beavis RC, Creasy D, Horn DM, Julian Jr RK, Seymour SL, Taylor CF, Vandenbrouck Y. Guidelines for reporting the use of mass spectrometry informatics in proteomics. Nat Biotechnol. 2008;26:862.

Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. Mol Cell Proteomics MCP. 2004;3:531–3.

Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20:1466–7.

da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. Nucleic Acids Res. 2006;34:D655–8.

Deutsch EW. The PeptideAtlas project. Methods Mol Biol. 2010;604:285–96.

Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R. A guided tour of the trans-proteomic pipeline. Proteomics. 2010;10:1150–9.

Deutsch EW, Chambers M, Neumann S, Levander F, Binz PA, Shofstahl J, Campbell DS, Mendoza L, Ovelleiro D, Helsens K, Martens L, Aebersold R, Moritz RL, Brusniak MY. TraML–a standard format for exchange of selected reaction monitoring transition lists. Mol Cell Proteomics MCP. 2012;11:R111.015040.

Domann PJ, Akashi S, Barbas C, Huang L, Lau W, Legido-Quigley C, McClean S, Neususs C, Perrett D, Quaglia M, Rapp E, Smallshaw L, Smith NW, Smyth WF, Taylor CF. Guidelines for reporting the use of capillary electrophoresis in proteomics. Nat Biotechnol. 2010;28:654–5.

Farrah T, Deutsch EW, Kreisberg R, Sun Z, Campbell DS, Mendoza L, Kusebauch U, Brusniak MY, Huttenhain R, Schiess R, Selevsek N, Aebersold R, Moritz RL. PASSEL: the PeptideAtlas SRMexperiment library. Proteomics. 2012;12:1170–5.

Garwood K, McLaughlin T, Garwood C, Joens S, Morrison N, Taylor CF, Carroll K, Evans C, Whetton AD, Hart S, Stead D, Yin Z, Brown AJ, Hesketh A, Chater K, Hansson L, Mewissen M, Ghazal P, Howard J, Lilley KS, Gaskell SJ, Brass A, Hubbard SJ, Oliver SG, Paton NW. PEDRo: a database for storing, searching and disseminating experimental proteomics data. BMC Genomics. 2004;5:68.

Gibson F, Anderson L, Babnigg G, Baker M, Berth M, Binz PA, Borthwick A, Cash P, Day BW, Friedman DB, Garland D, Gutstein HB, Hoogland C, Jones NA, Khan A, Klose J, Lamond AI, Lemkin PF, Lilley KS, Minden J, Morris NJ, Paton NW, Pisano MR, Prime JE, Rabilloud T, Stead DA, Taylor CF, Voshol H, Wipat A, Jones AR. Guidelines for reporting the use of gel electrophoresis in proteomics. Nat Biotechnol. 2008;26:863–4.

Gibson F, Hoogland C, Martinez-Bartolome S, Medina-Aunon JA, Albar JP, Babnigg G, Wipat A, Hermjakob H, Almeida JS, Stanislaus R, Paton NW, Jones AR. The gel electrophoresis markup language (GelML) from the Proteomics Standards Initiative. Proteomics. 2010;10:3073–81.

Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N, Cox J, Neumann S, Fan J, Reisinger F, Xu QW, Del Toro N, Perez-Riverol Y, Ghali F, Bandeira N, Xenarios I, Kohlbacher O, Vizcaino JA, Hermjakob H. The mzTab data exchange format: communicating MS-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics 2014. mcp.O113.036681.

Hoogland C, O'Gorman M, Bogard P, Gibson F, Berth M, Cockell SJ, Ekefjard A, Forsstrom-Olsson O, Kapferer A, Nilsson M, Martinez-Bartolome S, Albar JP, Echevarria-Zomeno S, Martinez-Gomariz M, Joets J, Binz PA, Taylor CF, Dowsey A, Jones AR. Guidelines for reporting the use of gel image informatics in proteomics. Nat Biotechnol. 2010;28:655–6.

Jiao X, Sherman BT, da Huang W, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012;28:1805–6.

Jones AR, Carroll K, Knight D, Maclellan K, Domann PJ, Legido-Quigley C, Huang L, Smallshaw L, Mirzaei H, Shofstahl J, Paton NW. Guidelines for reporting the use of column chromatography in proteomics. Nat Biotechnol. 2010;28:654.

Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard S, Selley J, Searle B, Shofstahl J, Seymour S, Julian R, Binz PA, Deutsch EW, Hermjakob H, Reisinger F, Griss J, Vizcaino JA, Chambers M, Pizarro A, Creasy D. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics MCP. 2012;11:M111.014381.

Kaiser J. Proteomics. Public-private group maps out initiatives. Science. 2002;296:827.

Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz PA, Deutsch EW. mzML-a community standard for mass spectrometry data. Mol Cell Proteomics MCP. 2011;10:R110.000133.

Martinez-Bartolome S, Deutsch EW, Binz PA, Jones AR, Eisenacher M, Mayer G, Campos A, Canals F, Bech-Serra JJ, Carrascal M, Gay M, Paradela A, Navajas R, Marcilla M, Hernaez ML, Gutierrez-Blazquez MD, Velarde LF, Aloria K, Beaskoetxea J, Medina-Aunon JA, Albar JP. Guidelines for reporting quantitative mass spectrometry based experiments in proteomics. J Proteomics. 2013;95:84–8.

Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tarraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, Garcia F, Marba M, Montaner D, Dopazo J. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 2010;38:W210–3.

Medina-Aunon JA, Paradela A, Macht M, Thiele H, Corthals G, Albar JP. Protein information and knowledge extractor: discovering biological information from proteomics data. Proteomics. 2010;10:3262–71.

Orchard S, Hermjakob H, Apweiler R. The proteomics standards initiative. Proteomics. 2003;3:1374–6.

Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol. 2007;25:894–8.

Orchard S, Binz PA, Borchers C, Gilson MK, Jones AR, Nicola G, Vizcaino JA, Deutsch EW, Hermjakob H. Ten years of standardizing proteomic data: a report on the HUPO-PSI Spring Workshop: April 12-14th, 2012, San Diego, USA. Proteomics. 2012;12:2767–72.

Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol. 2004;22:1459–66.

Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res. 2012;40:W478–83.

Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, Mohammed S, Deery MJ, Howard JA, Dunkley T, Aebersold R, Kell DB, Lilley KS, Roepstorff P, Yates 3rd JR, Brass A, Brown AJ, Cash P, Gaskell SJ, Hubbard SJ, Oliver SG. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nat Biotechnol. 2003;21:247–54.

Taylor CF, Paton NW, Lilley KS, Binz PA, Julian Jr RK, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates 3rd JR, Hermjakob H. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007;25:887–93.

Taylor CF, Binz PA, Aebersold R, Affolter M, Barkovich R, Deutsch EW, Horn DM, Huhmer A, Kussmann M, Lilley K, Macht M, Mann M, Muller D, Neubert TA, Nickson J, Patterson SD, Raso R, Resing K, Seymour SL, Tsugita A, Xenarios I, Zeng R, Julian Jr RK. Guidelines for reporting the use of mass spectrometry in proteomics. Nat Biotechnol. 2008;26:860–1.

Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, O'Kelly G, Schoenegger A, Ovelleiro D, Perez-Riverol Y, Reisinger F, Rios D, Wang R, Hermjakob H. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 2013;41:D1063–9.

Walzer M, Qi D, Mayer G, Uszkoreit J, Eisenacher M, Sachsenberg T, Gonzalez-Galarza FF, Fan J, Bessant C, Deutsch EW, Reisinger F, Vizcaino JA, Medina-Aunon JA, Albar JP, Kohlbacher O, Jones AR. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. Mol Cell Proteomics MCP. 2013;12:2332–40.

Chapter 5
Mass Spectrometry-Based Protein Sequencing Platforms

Toshihide Nishimura and Hiromasa Tojo

5.1 Introduction

The amino-terminal sequencing method was first introduced by the stepwise degradation of peptides applied in 1930 by Abderhalden and Brockmann (1930), which was improved by Pehr Edman (1949) in 1949. Until mid 1980s, the "Edman degradation" procedure had been used to determine the sequence of the N-terminal 30–40 amino acid residues in a protein routinely. The term "proteome", linguistically equivalent to the concept of genome, was coined in 1994 to describe the complete set of proteins that is expressed according to the genome information and modified following expression in a lifetime of cells (Nature 1999). A simultaneous, high-throughput sequencing for more than thousands of peptides/proteins in complex biological samples had not been possible until both soft ionizations for biological molecules (MALDI, ESI etc.) (Tanaka et al. 1988; Karas and Hillenkamp 1988; Yamashita and Fenn 1984; Fenn et al. 1989) and mass spectrometry (MS)-driven sequencing technologies have been developed. This MS-based proteomics has dramatically revolutionized the sequencing and identification of proteins/peptides in complex biological samples. Clinical proteomics is nowadays a science to understand dynamic protein-centric biomolecular networks spatially and temporally in diseased cells and organs.

T. Nishimura (✉)
First Department of Surgery, Tokyo Medical University, Tokyo, Japan

R&D, Biosys Technologies, Tokyo, Japan
e-mail: t-nisimura-tmu@hotmail.co.jp

H. Tojo
Department of Biophysics and Biochemistry, Graduate School of Medicine,
Osaka University, Suita, Japan

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_5

5.2 MS-Based Sequencing of Polypeptides

In the tandem MS (MS/MS) (McLafferty 1983) and MS^n experiments, the first mass analyzer selectively passes ions of interest into another reaction region where excitation and dissociation take place, and then the second mass analyzer records the *m/z* values of the fragment ions. Precursor ions are most commonly excited by energetic collisions with non-reactive gas such as Ar and He (collision-induced (activated) dissociation) (CID or CAD). The observed fragmentation pattern depends on various parameters including the amino-acid composition and size of the peptide, time-scale of the instrument, the charge state of the ions, etc. Under the most usual low-energy collision conditions, peptide precursor ions fragment along the backbone at the peptide bonds, forming structurally informative sequence ions and also less useful non-sequence ions formed by losing water, ammonia, etc. The nomenclature of possible sequence ions of peptides was proposed by Roepstorff and Fohlman (1984) and by Biemann (1988), Biemann and Papayannopoulos (1994) (Fig. 5.1). The cleavage at the Cα—C, C—N, and N—Cα bonds of a peptide linkage can form the respective charged N-terminal fragments *a*, *b* and *c* ions, and the corresponding charged C-terminal fragments *x*, *y* and z ions. The numbering indicates which peptide bond is cleaved, counting from the N- to C-terminus for the former ions and in reverse order for the latter ions, and corresponds to the number of amino acids in the fragment ion.

5.2.1 Collision-Induced Dissociation (CID) (Paizs and Suhai 2005)

The most representative model describing how protonated peptides dissociate upon excitation will be the mobile proton model, which has emerged as a result of a large number of experimental and theoretical studies (Paizs and Suhai 2005; Wysocki et al.

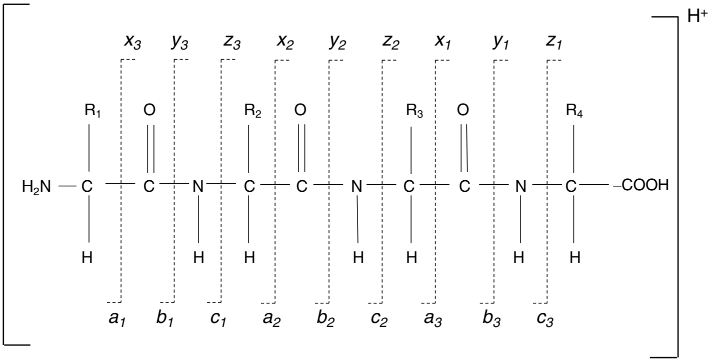


Fig. 5.1 The nomenclature of structurally informative sequence ions, the N-terminal a, b, and c, and the C-terminal x, y, and z ions. Non-sequence ions are also formed by neutral loss of water, ammonia, etc

2000; Harrison and Yalcin 1997; Summerfield et al. 1997; Tang et al. 1993). Protonated peptide ions in low-energy CIDs mostly undergo charge-directed fragmentations. Peptides can be protonated at various sites such as N-terminal amino group, amide oxygens and nitrogens, and side chain functional groups to form peptide ion isomers with a proton at different sites. There are two major classes of these peptide ions: In the first group, one or more of the protonation sites is energetically more favored than the others, and so the attached proton(s) are sequestrated there. For example, in a singly protonated tryptic peptide containing arginine at the C-terminus its charge is localized at the arginine side chain. For the second group, multiple protonation sites are accessible in a relatively narrow energy range, forming doubly charged tryptic peptides, which are practically useful in peptide identification.

Energetically less favored protonation sites of the ions can be more populated with increasing their internal energy upon excitation. Then the reaction channel of a charge-remote fragmentation can be opened when a proton is hardly sequestrated. The recently refined the 'mobile proton model' involving mechanistic considerations allows qualitative understanding of the fragmentation behavior of peptides containing a few arginines and/or acidic residues and/or protons, (Tsaprailis et al. 1999) and has been applied to interpret the charge-state dependent fragmentation of gaseous protein ions (Engel et al. 2002). It can be noted that non-selective fragmentation is observed when the number of ionizing protons is larger than the number of arginines, and that selective fragmentation via the aspartic acid effect is expected when the number of arginines is larger or equal to the number of ionizing protons (Csonka et al. 2000, 2001; Paizs and Suhai 2001a, b; Polce et al. 2000).

5.2.1.1 Charge-Directed Fragmentation Pathways

Protonated amide bonds can be cleaved either by rearrangement-type reactions or by direct bond cleavage. In low-energy collisions, a protonated peptide ion mostly dissociates via a rearrangement reaction prior to its backbone cleavage. Therefore, such a fragment ion is mostly formed due to non-specific fragmentation, wherein dissociation is induced by attacking the carbon center of the protonated amide bond either by the N-terminal neighbor amide oxygen ($b_x - y_z$ fragmentation pathway) or by the nitrogen of the N-terminal amino group (diketopiperazine$-y_{N\text{-}2}$ fragmentation pathway). Figure 5.2 schematically shows those two fragmentation pathways.

In the b_x – y_z pathway shown in Fig. 5.2a, fragmentation of the proton-bound intermediate complex (*i*) would be governed by the thermochemical quantity of the neutral fragments, oxazolone derivatives which are the neutral counterparts of the y_z ions. In such cases, the abundance ratio between the b_x and y_z ions can follow the linear free-energy relationship (Harrison 1999; Paizs and Suhai 2002, 2004; Morgan and Bursey 1994):

$$ln\left(\frac{b_x}{y_z}\right) = \frac{\left(PA_{N-term} - PA_{C-term}\right)}{RT_e} \tag{5.1}$$

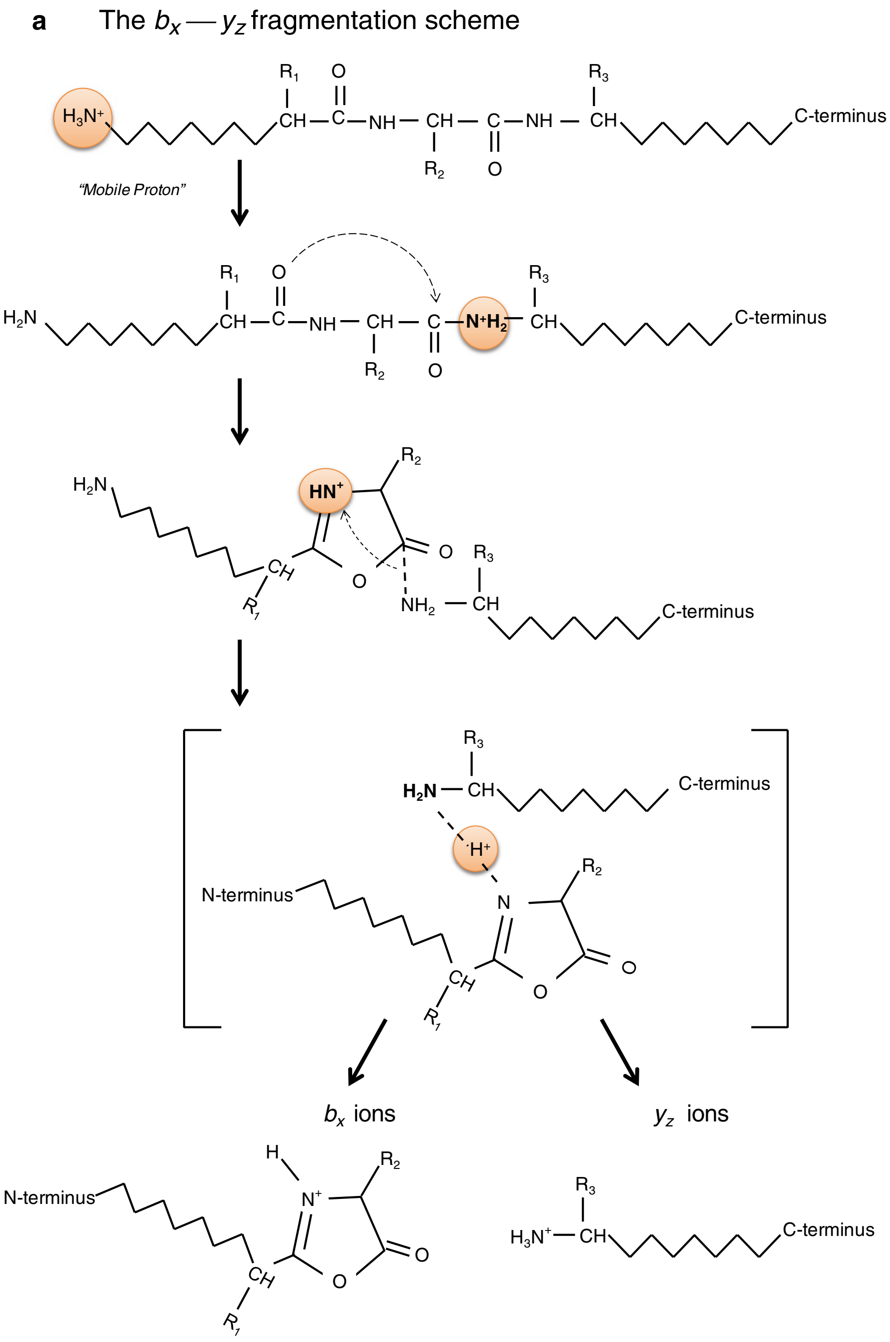


Fig. 5.2 The two typical fragmentation schemes following proton migration on a protonated peptide ion: (**a**) the $b_x - y_z$ and (**b**) the *diketopiperazine* – $y_{N\text{-}2}$ pathways (Modified from Paizs and Suhai 2005)

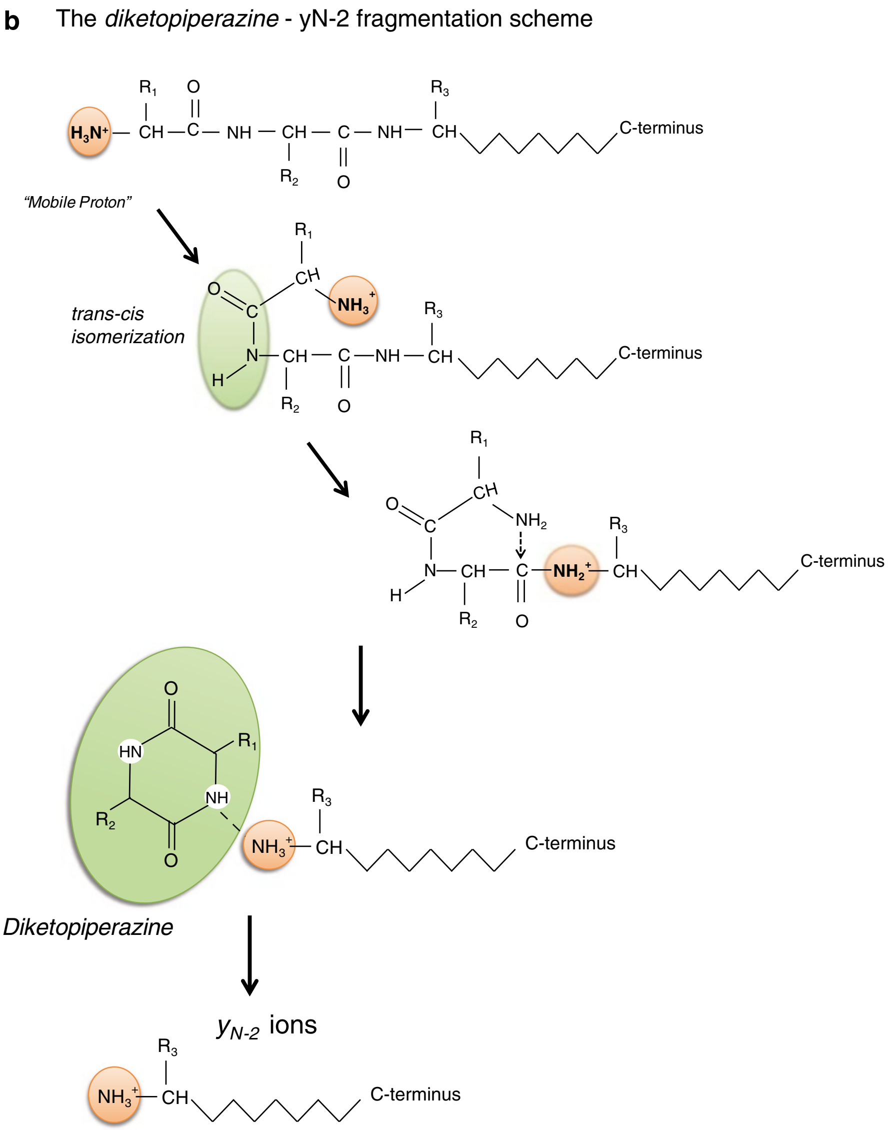


Fig. 5.2 (continued)

Herein, $PA_{N\text{-}term}$ and $PA_{C\text{-}term}$ are the proton affinities (*PA*) of the neutral fragments in the $b_x - y_z$ pathway, and T_e the effective temperature. Figure 5.3 gives an example, in which the linear relationship between log (y_1/b_2) and *PA* is shown for the C-terminal amino acid residue for protonated tripeptides of the series, H-Gly–Gly-Xxx-OH (Harrison 1999).

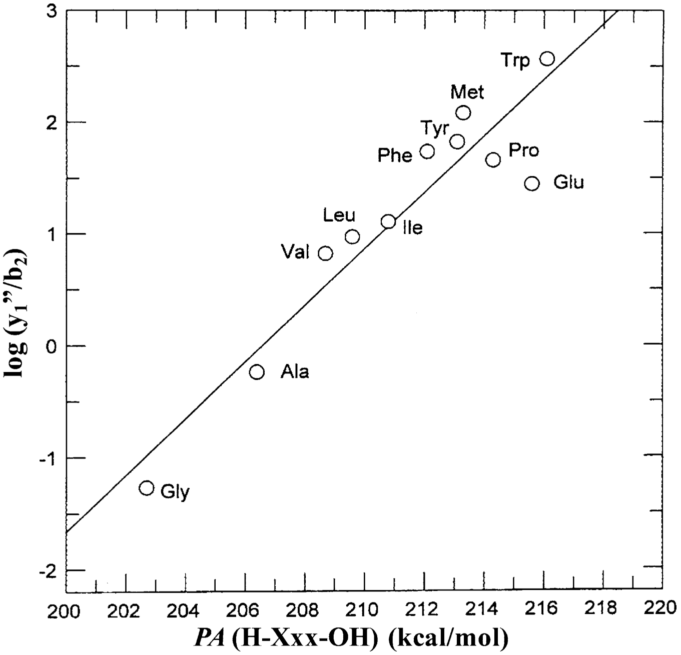


Fig. 5.3 A linear free energy relationship between the intensity ratio y_1"/b_2 and proton affinity of amino acids, *PA*(H-Xxx-OH). The intensity data were taken from 10-eV CID of protonated H-Gly-Gly-Xxx-OH (Adapted with modification from Harrison 1999)

Most protonated proline containing peptides are characteristic to their MS/MS spectra fragmentation patterns distinct by abundant *y* ions with N-terminal to the proline residues. Although Schwartz and Bursey (1992) interpreted the proline effect on the basis of the high PA of Pro by CID studies, it seems that the proline residue has no specific effect on the cleavage of the N-terminal amide bond of protonated peptides and that such amide bonds can be cleaved in line with the rules of the $a_1 - y_z$ pathway, where the formation of *y* ions containing Pro at the N-terminal would be one of the low- energy processes (Vaisar and Urban 1996). However, it is true that Pro has a rather specific effect when located at the C-terminal.

5.2.1.2 Charge-Remote Fragmentations

Selective cleavages at acidic residues including Asp, Glu, and Cys have been studied for a large number of different peptides, (Tsaprailis et al. 1999) and the situations are illustrated in Fig. 5.4: (a) Nonselective cleavages along the peptide backbone may occur when the number of ionizing protons exceeds the number of arginine residues, and (b) Selective cleavages adjacent to the acidic residues may become predominant when the number of ionizing protons equals the number of arginine residues. Selective fragmentation can be interpreted by the effective

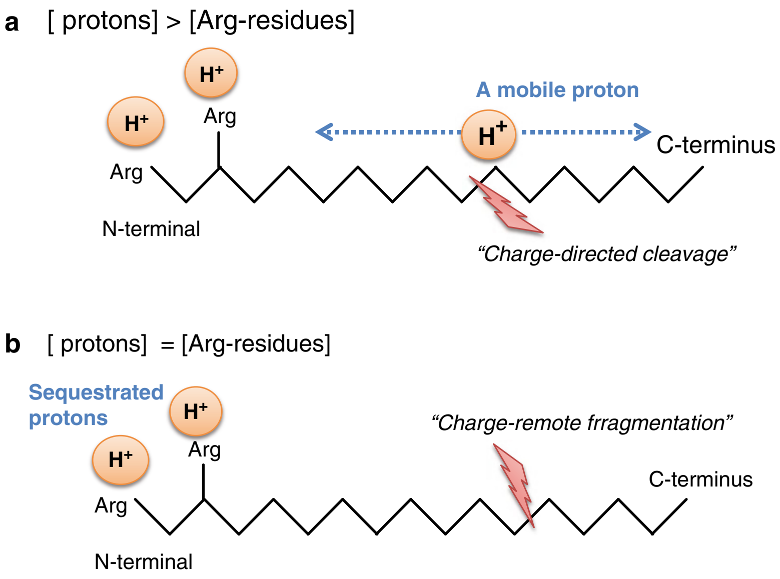

a [protons] > [Arg-residues]
H+
H+
Arg
Arg
A mobile proton
H+
C-terminus
N-terminal
"Charge-directed cleavage"
b [protons] = [Arg-residues]
Sequestrated protons
H+
H+
Arg
Arg
"Charge-remote frragmentation"
C-terminus
N-terminal

Fig. 5.4 (**a**) Nonselective cleavages along the peptide backbone via charge directed fragmentation when [protons] > [arginine residues], and (**b**) a selective charge-remote fragmentation via adjacent to the acidic residues when [protons] = [arginine residues]

sequestration of the added proton(s) by the side chain of the arginine residue(s). Since the proton sequestrated peptide ions are energetically stable among other states, the reaction channel for charge-remote dissociation opens when it is energetically available.

Many of the protonated peptides containing aspartic or glutamic acid residues show distinct fragmentation behavior producing abundant *b* ions C-terminal to these residues (*Aspartic acid effect*). Facile cleavages at Asp–Pro peptide bonds studied by MALDI-TOFMS (YuW et al. 1993) have suggested that the Asp-Pro peptide bond is more labile than the other peptide bonds regardless of the size of peptides. It has been demonstrated that the esterification of the COOH group of the Asp side chain shuts down dominant formation of the *b* ion C-terminal to Asp, suggesting for the role of the side chain in catalyzing the fragmentation intramolecularly. The CID experiments on peptides containing a fixed charge and Asp residues (Gu et al. 2000) have proven that the charge remote fragmentation is responsible for the aspartic acid effect, in which the number of added protons and the number of arginine residues are equal.

Since methionine oxidation frequently occurs during sample preparation and storage, a charge-remote fragmentation of singly protonated tryptic peptide ions leads to the 2-cis-elimination of CH_3SOH from the neutral side chain (*Loss of methane sulfenic acid*) (O'Hair and Reid 1999). The labeling and MS^3 experiments suggested that this elimination reaction is also competitive with the charge directed fragmentations of singly charged tryptic peptide ions containing oxidized methionine.

5.2.2 Electron-Transfer Dissociation (ETD)

Zubarev et al. (1998) and Kruger et al. (1999) introduced a novel fragmentation technique, electron capture dissociation (ECD) that is induced by the capture of low-energy thermal electrons introduced directly in the superconducting magnet portion of Fourier Transform Ion Cyclotron Resonance (FTICR) mass analyzers by multiply- protonated peptide ions. Therefore, its applicability is so far restricted to high-end mass spectrometers equipped with superconducting magnet fields. Recently, ECD-like fragmentation has been found to be induced by electron transfer on the reaction of peptide cations with reagent radical anions. This electron transfer dissociation (ETD) technique can be implemented on ion trap mass spectrometers available commercially (Coon et al. 2004; Syka et al. 2004).

The electron transfer dissociation (ETD) takes place by the reaction of multiply protonated peptide cations, [MH_n^{n+}], with small-molecule anions, [$A^{\bullet-}$]. The first step of these reactions is the transfer of an electron from the anion to the peptide cation.

$$MH_n^{n+} + A^{\bullet-} \rightarrow MH_n^{\bullet(n-1)+} + A \tag{5.2}$$

Such a charge-reduced multiply protonated peptide ion [$MH_n^{\bullet(n-1)+}$] is an odd-electron cation that undergoes free-radical-driven cleavage, which results in dissociation of the N-C$_\alpha$ bond, typically generating even electron *c*-type and odd electron •*z*-type product ions.

5.2.2.1 Direct Dissociation-Like Cleavage of N-C$_\alpha$ Bond

It is desirable in proteomic sequencing that peptide backbone bonds can randomly be cleaved regardless of peptide length, *z*, *m*/z, and amino acid composition or order, and also with the presence of posttranslational modifications (PTMs). CID occurs typically in a time scale slower than *nano*-seconds (10^{-9}–10^{-6} s) during which initial excitation energies can redistribute throughout the peptide cations to induce cleavage of the weakest bonds, such as phosphorylation, glycosylation, and sulfonation. In phosphorylated peptide ions, their CIDs result readily in displacement of phosphoric acid (H_3PO_4) by leaving the backbone bonds intact (Fig. 5.5). In contrast, ETD fragmentation can allow every possible backbone fragment without losing PTMs, and that ETD provides sufficient backbone cleavage for identification of both sequence and site localization of PTMs.

The N–C$_\alpha$ bond cleavage by ETD occurs faster than *pico*-seconds (10^{-12} s) via a direct dissociation-like process that does not involve the redistribution of intramolecular vibrational energies whereas protons are mostly mobile before electron transfer reactions. In principle, all recombination energy is thus concentrated in just a few bonds around the cleavage site, which would result in backbone fragment ions conserving PTMs such as phosphorylation. ETD can be used for *de-novo* sequencing and characterization of PTMs and gas-phase structures, which have been difficult for CID to address.

Syrstad and Tureček (2001) investigated the dissociation energetics of charge reduced peptide ions in electron capture dissociation (ECD) by *ab initio* and density

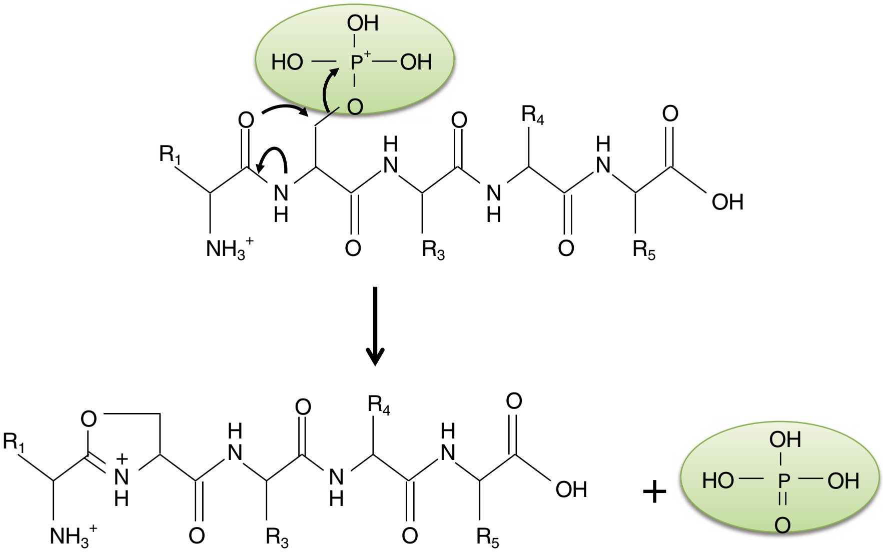


Fig. 5.5 A weak phospho-monoester bond scission due to internal energy redistribution in CID

functional theory calculations, and concluded that aminoketyl radicals form as key transient intermediates. There are several long-lived excited states in which the amide carbonyl group incurs a proton affinity exceeding 1200 kJ/mol as superbases, which is more than sufficient to abstract a proton from any protonated amino acid residue. Those long-lived excited state intermediates would be involved in proton transfer, and then forming charge reduced peptides with a labile aminoketyl radical that can undergo facile cleavage of the adjacent N-C_α bond (Fig. 5.6).

The experimental and theoretical studies by Chung and Tureček (2011) concluded that aminoketyl radicals play a key role in N-C_α bond cleavage of odd-electron peptide cation-radicals, and that there are two distinct types referred to as the "proper" and "improper" aminoketyl radicals (Chung and Tureček 2011). The "proper" aminoketyl radicals have pyramidized structures, in which the high spin density is localized, and from which the direct N-C_α bond cleavage occurs. In contrast, the "improper" ones have near planar $-C_\alpha C(OH)NH-$ groups, in which the spin density is delocalized over several atoms adjacent to the aminoketyl moiety, and need higher transition energies for N-C_α bond cleavage and are energetically favor to undergo H-atom transfers resulting in side-chain losses (Chung and Tureček 2010; Tureček and Syrstad 2003; Yao et al. 2007; Gilbert and Smith 1990).

5.2.2.2 Decision Tree (DT)-Driven Shotgun Sequencing

In Shotgun proteomic experiments, highly complex samples typically containing hundreds of thousands of tryptic peptides are subjected to separation by nano flow-rate liquid-chromatography (nanoLC) followed by mass spectrometric sequencing

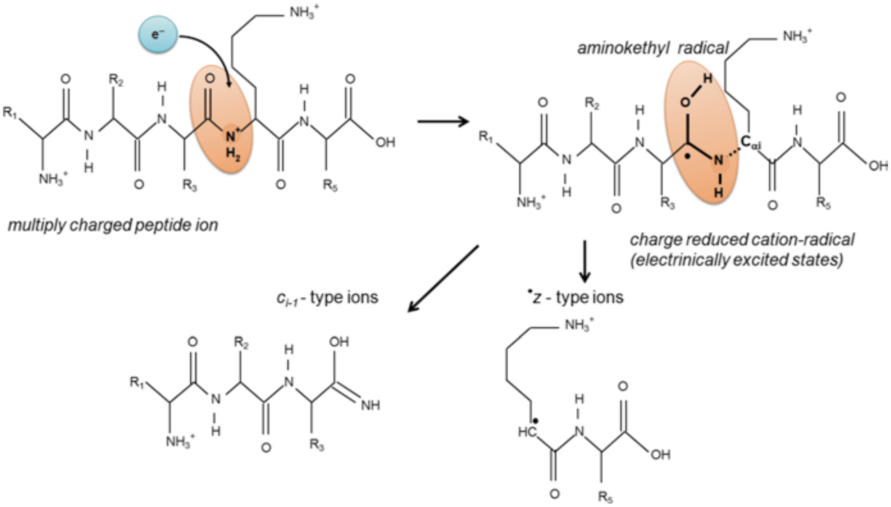
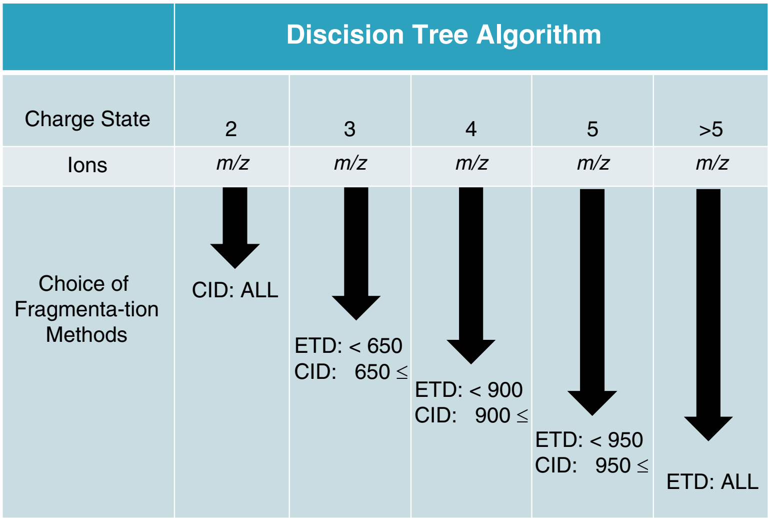

e⁻
multiply charged peptide ion
aminokethyl radical
charge reduced cation-radical
(electrinically excited states)
c_{i-1} - type ions
•z - type ions

Fig. 5.6 The facile N-C_α bond cleavage via a charge reduced cation radical intermediate containing a labile aminoketyl radical in ETD, which results in c and •z ions

of their MS/MS spectra. Since ETD has the potential usefulness in sequencing peptides with intact PTMs and in improving the peptide sequence coverage, its applicability to clinical proteomic experiments of large-scale is now worthwhile to be investigated. Coon et al. (Swaney et al. 2008; Coon 2009; Swaney et al. 2007) demonstrated an effective peptide sequencing and protein identification for complex peptide mixtures by optimal multiple dissociation methods, CID and ETD, following the data dependent decision tree (DT) (Fig. 5.7). Their large-scale proteome analysis of Saccharomyces cerevisiae and human embryonic stem cells (hES) with the DT algorithm netted 53,055 peptide identifications whereas the peptide identifications were 38,293 (CID) or 39,507 (ETD) by the individual methods alone. Similarly, the application of the DT method to phosphoproteomics yielded 7,422 vs. either 2,801 (CAD) or 5,874 (ETD) phosphopeptides.

5.2.2.3 Peptide Sequencing by Dual Fragmentation (Frese et al. 2012)

The ETD is in principle the N−C_α backbone cleavage of odd electron peptide cations, which results primarily in c- and •z-type product ions. This electron-driven fragmentation is fundamentally different from CID. It has been noted that highly-protonated peptide ions are favorable for ETD-based peptide sequencing, and that ETD is most suitable for longer and basic peptides with more than triply-charge states. In the ETD process, the charge-reduction by electron transfer results immediately in the formation of an intermediate radical cation, $[MH_n]^{(n-1)+\bullet}$. It has been frequently observed that this non-dissociative, charge reduced precursor and the



	Discision Tree Algorithm				
Charge State	2	3	4	5	>5
Ions	*m/z*	*m/z*	*m/z*	*m/z*	*m/z*
Choice of Fragmenta-tion Methods	CID: ALL	ETD: < 650 CID: 650 ≤	ETD: < 900 CID: 900 ≤	ETD: < 950 CID: 950 ≤	ETD: ALL

Fig. 5.7 The schematic diagram of the DT algorithm; applies to The precursor ions in the indicated *m/z* value range are selected for MS/MS interrogation indicated (Adapted with modification from Coon 2009)

unreacted precursor ion remain dominant in MS/MS spectra. Coon et al. (Swaney et al. 2007) introduced the dual fragmentation method that utilizes low-energy CID of the charge reduced precursor subsequent to the electron-transfer process (ETcaD). This method enhances the fragmentation efficiency and results in increased formation of c- and z-ions, especially of doubly charged peptides.

Frese et al. (2012) have also introduced the dual fragmentation method abbreviated as EThcD by combining ETD followed by higher-energy collision dissociation (HCD). Figure 5.8 compares tandem mass spectra taken under various fragmentation conditions: ETD, ETcaD, HCD, and EThcD. In EThcD, after an initial electron-transfer dissociation, all resulted ions including the unreacted precursor ions are subjected further to high-energy collision induced dissociation which yields *b/y*- and *c/z*-type fragment ions in a single spectrum. EThcD thus provides rich sequence information with a substantially increase of the peptide sequence coverage.

5.3 LC/MS-Based Shotgun Cancer Proteomics

Shotgun proteomics has been over decades the most powerful in characterizing complex proteomes of clinical samples (Steen and Mann 2004; Marcotte 2007; Yates et al. 2009; Marko-Varga et al. 2007). Its analytical platform is comprised of a single or multidimensional (MD) nano-scale liquid chromatography-tandem mass spectrometry (LC-MS/MS) followed by database search and protein inference

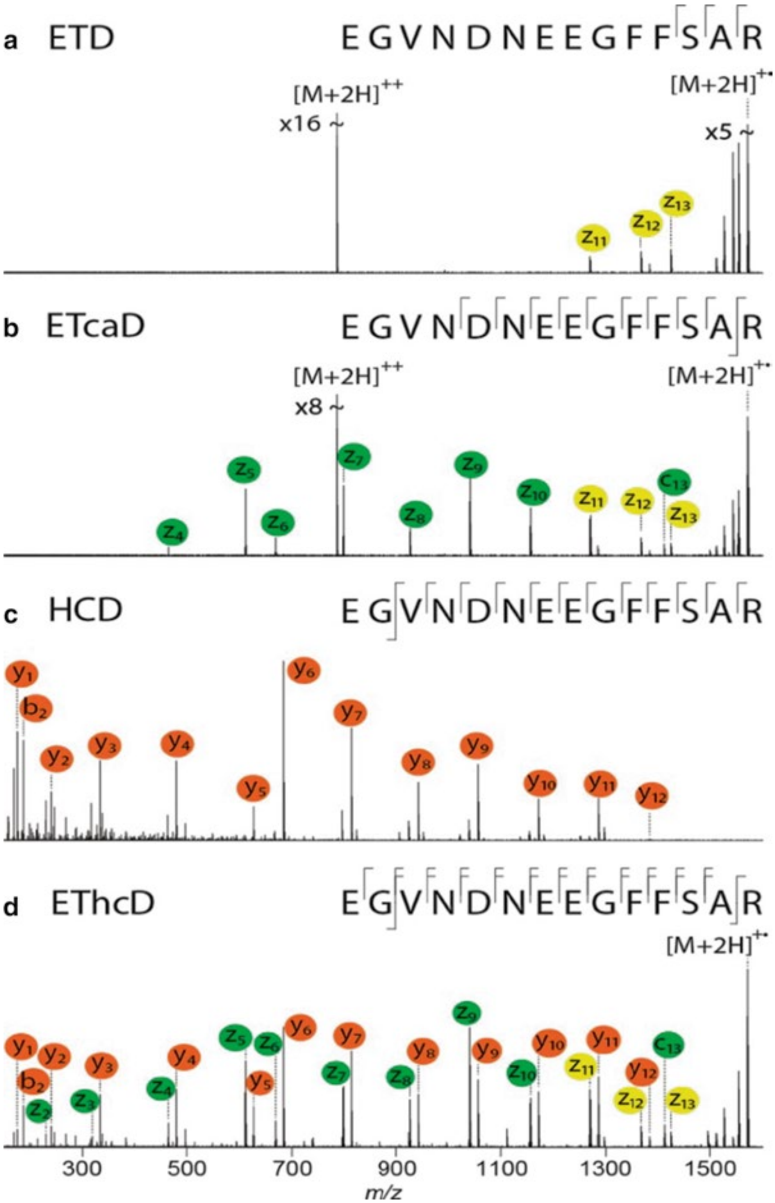


Fig. 5.8 Tandem mass spectra of the doubly-charged peptide EGVNDNEEGFFSAR (i.e., Glufib) ions by (**a**) ETD, (**b**) ETcaD, (**c**) HCD or (**d**) EThcD. Colors represent the species of the fragment ions: 🟡, c/z-ions generated by ETD; 🟢, c/z-ions derived from CID of the charge reduced precursor; 🔴, b/y ions derived from HCD of the precursor (Adapted from Frese et al. 2012)

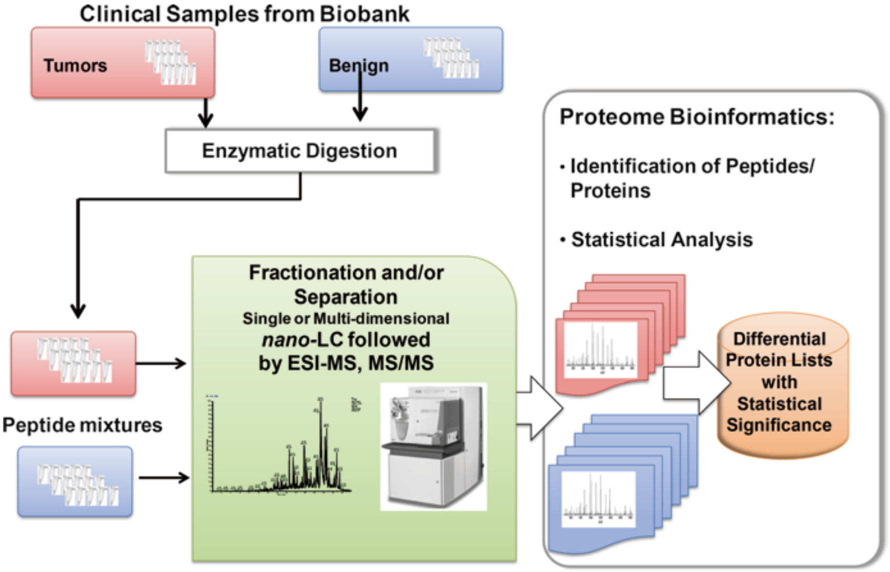

Clinical Samples from Biobank
Tumors
Benign
Enzymatic Digestion
Peptide mixtures
Fractionation and/or Separation
Single or Multi-dimensional
nano-LC followed by ESI-MS, MS/MS
Proteome Bioinformatics:
• Identification of Peptides/ Proteins
• Statistical Analysis
Differential Protein Lists with Statistical Significance

Fig. 5.9 A Shotgun-proteomic analysis platform comprised of a single or multidimensional (MD) *nano*-scale liquid chromatography-tandem mass spectrometry (LC-MS & MS/MS), database search and protein inference algorithms, and a variety of statistical evaluation of obtained proteomic data

algorithms, shotgun proteomics platforms surpass other MS-based proteomics systems in number and diversity of proteins identified and in dynamic range for detection. A brief illustration of the shotgun proteomic analytical platform is shown in Fig. 5.9. Nevertheless, shotgun proteomics using LC-MS/MS is essentially a sampling technique, in which probability of detection is a function of protein abundance and quantitation is assessed by counting the numbers of spectra that maps to proteins identified. However, random sampling of medium to low abundance proteins in shotgun analyses means that multiple replicate analyses are needed to establish the composition of complex proteomes. Because shotgun analyses can represent complex proteomes in considerable depth, a key question is whether comparison of shotgun proteome inventories can reveal molecular characteristics of biologically distinct phenotypes.

5.3.1 Proteolysis Prior to Sequencing Proteins

In sequencing proteins it is preferable that amide linkages within precursor ions are randomly protonated so that a variety of informative fragment ions are produced by CID and/or ETD. Specific tryptic digestion on the C-terminal side of lysine (K) and arginine (R) residues is suitable for promoting random protonation because tryptic peptides do not contain any internal basic amino acid which has so high a proton affinity to block random protonation. A drawback of trypsin digestion is that it generates relatively shorter peptides. Although numerous other enzymes are available and produce longer-size peptides, however, many of these peptides contain internal basic residues which sequester charge and prevent random backbone cleavages.

5.3.2 Sample Preparations of Clinical Specimens

Millions of clinical samples are obtained everyday for use in diagnostic tests that support clinical decision-making. Clinical samples (tissue, biopsy, blood, etc.) can also be archived into repositories for use in future studies investigating the etiology of diseases using omics approaches. Therefore, the infrastructure buildup of standardized biobanking is increasingly needed within the clinical omics community because the samples themselves have intrinsic values as to determine outcomes of clinical trials (Végvári et al. 2011a, b; Marko-Varga 2011; Marko-Varga et al. 2011; Malm et al. 2012; LaBaer 2012). The samples can be retrieved from pathology laboratories with the approval from Ethical Committee of Medical Institutes and Hospitals. There are many kinds of disease specimens such as frozen and formalin-fixed paraffin-embedded (FFPE) tissues, biopsies, and body fluids containing blood, serum, plasma and urine, interstitial fluid, cyst, ascites fluid, pancreatic juice, etc. Here, we briefly describes sample preparations for plasmas and FFPE tissues, representive types of clinical specimens.

5.3.2.1 Plasma Samples

Blood is representative among body fluids and is of less-invasive clinical specimens, and has been preferably utilized for most of clinical proteomic studies (Anderson 2010a). The Human Proteome Organization (HUPO; www.HUPO.org) recommends plasma rather than serum with ethylene diamine tetraacetic acid (EDTA) as an anticoagulant (Omenn et al. 2005). The abundant proteins in plasma are usually depleted prior to analysis (Omenn et al. 2005). Acceptance of protein annotation, i.e., accepted protein identities (Omenn et al. 2005; Martens et al. 2005) should use standard criteria. These include having two identified peptide sequences from each protein, both with a statistical significance score high enough to ensure a correct sequence confirmation when compared with the corresponding gene sequence entity. For a large number of plasma samples, it is critical to optimize preparation protocols to ensure stability in all procedures, in addition to both quality assessment of sample preparations and randomization of the processed samples.

The abundance range of plasma proteomes extends at least 10–11 orders of magnitude with albumin and hemoglobin the most abundant and with interleukins the least abundant (at *pg/ml*) (Anderson and Anderson 2002). Since this wide dynamic range of protein concentrations presents a barrier to detection of medium and low abundance proteins in proteomic analyses, targeted depletion of abundant plasma proteins with antibody columns has been employed both to increase the depth of proteome identifications to increase sensitivity for targeted analyses of specific proteins. Immobilized antibody columns and removal kits are commercially available. For example, Multiple Affinity Removal System™ (MARS) columns (4.6 × 100 mm) designed to deplete 7 abundant proteins (albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, fibrinogen) (MARS-7) or to deplete 14 abundant proteins (albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, fibrinogen, alpha2-macroglobulin, alpha1-acid glycoprotein, IgM, apolipoprotein AI, apolipoprotein AII, complement C3, and transthyretin) (MARS-14). A SuperMix column (GeneWay, Inc.) could immune-depletes 81 most abundant proteins. A global LC-MS/MS analysis of multiply immune-depleted plasma can provide much-improved protein profiling of patient's blood (Kovács et al. 2011).

It has to be, however, noted that those antibody-based depletion systems are not standardized, and so variability in immune responses is carried over from sample preparation, which introduces a definite bias in discovering biomarkers factors with which distinguish sample groups as treatment and disease development. Simultaneously, an intrinsic question is thrown what about co-immuno-depleted proteins with the most abundant proteins. Albumin is the most abundant proteins and well known to interact with other molecules by its nature. Therefore, albuminome (all proteins in albumin fractions obtained with depletion treatment) is a subject of several studies that would provide an insight into the composition of co-depleted subproteomes (Gundry et al. 2007).

5.3.2.2 FFPE Tissue Specimens

In hospitals and medical institutes, tumor tissues obtained by surgical resection are typically fixed in 4 % paraformaldehyde and routinely processed for paraffin sectioning. Cancerous lesions are identified on serial sections stained with hematoxylin and eosin (HE-staining). For shotgun proteomic analysis, 10-μm sections prepared from the same tissue block are attached to glass or special slides. FFPE tissues on slides are de-paraffinized three times with xylene for 5 min, rehydrated with graded ethanol solutions and distilled water, and then stained with hematoxylin. Stained, uncovered slides are air-dried and typically more than 30,000 cancerous cells (8 mm^2) are collected by laser microdissection (LMD) techniques (Prieto et al. 2005; Hood et al. 2005; Kawamura et al. 2010; Nomura et al. 2011) using instruments such as Leica LMD7000 (Leica Microsystems GmbH, Germany), Axio Observer (Carl Zeiss Microscopy GmbH, Germany) or others.

Figure 5.10 shows an example of the images taken using a Leica LMD7000 (Leica Microsystems GmbH, Germany) before and after laser microdissection to collect about 20,000 prostate cancer cells per a tissue section on the dedicated DIRECTOR slide (Expression Pathology Inc., USA) that has a special coating to help harvest excised micro-tissues into a plastic tube. Proteins/peptides from dissected cells can be extracted by following several protocols (Prieto et al. 2005;

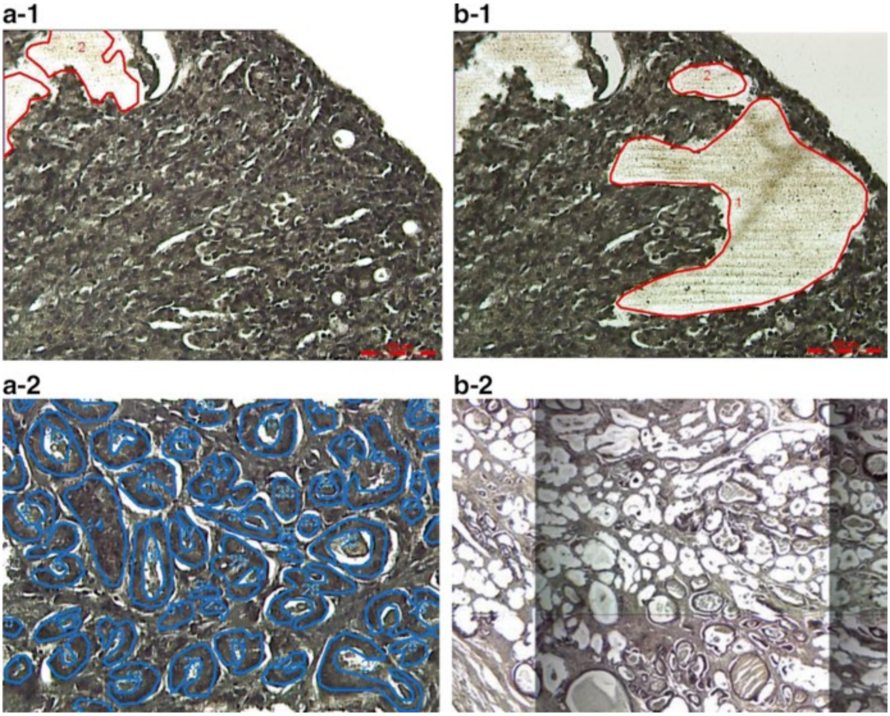


Fig. 5.10 Laser micro-dissection sampling from prostate cancer FFPE sections. (**a**) before and (**b**) after collecting ca. 20,000 prostate cancer cells each from FFPE tissues of the high and low Gleason scores (1, *top* and 2, *bottom*, respectively)

Wisniewski et al. 2011). For example, according to the protocol of a Liquid Tissue™ MS Protein Prep kit (Expression Pathology), (Prieto et al. 2005) the cellular material, suspended in the Liquid Tissue buffer, is incubated (95 °C for 90 min), then cool on ice (3 min), and then is enzymatically digested, followed by reduction and alkylation. The Liquid Tissue digests can be stored at −20 °C until proteomic analysis.

5.3.3 Differential Protein Expression Analysis

A group comparison of shotgun proteomic LC-MS data is often the first step to find out identified proteins unique to either disease or healthy states. In clinical proteome studies label-free semi-quantification methods have been preferably adopted; there are two label-free methods: (1) peak intensity-based methods and (2) spectral count-based methods. The latter approaches will be described in Sect. 5.4.

The peak intensity-based approaches have been so far popular in various fields, and would be believed to be more accurate than those of spectral counting. However, ion intensities of peptides are considerably affected by both their ionization efficiencies and ion suppression by co-eluting abundant ions, which might make the comparison of signal intensities challenging. In particular, the missing signals or identification of a given peptide from run to run is always found with typical frequencies ranging from 20 to 50 % in proteomics data sets, which causes the greatest bias in the analysis outcome. Accordingly, comparison analysis has been carried out on complete data sets, either excluding the missing values (Hill et al. 2008; Oberg et al. 2008) or filling in the missing values using standard imputation routines like k-nearest neighbors (KNN) (Troyanskaya et al. 2001). In these situations, excluded proteins with missing signals will inherently create a bias towards high-abundant proteins. Low-abundant proteins are often informative and functionally important when comparing disease states. In particular, intensity-based approaches are less amenable to identifying ‘one-state’ peptides that are present in one group, but not in another.

Differential expression methods using spectral counts are receiving an increasing acknowledgement (summarized briefly in Table 5.1). Such methods include NSAF (normalized spectral abundance factor) (Zybailov et al. 2006), PLGEM (power law global error model), (Pavelka et al. 2008) and SI_N (normalized spectral index) (Griffin et al. 2010). The recent spectral count-based methods such as SpI (spectral index) (Fu et al. 2008), Qspec using a Bayesian model (Choi et al. 2008) and Hybrid (Booth et al. 2011) take into account non-normal distribution and limited replicates

Table 5.1 A list of the differential expression methods based on spectral counting approaches

Modern differential expression methods					
			Consideration in methods		
Abbreviation	Method name	References	Non-normal distribution	Replicates and samples sizes	Applicability to multivariate tests
NSAF	Normalized Spectral Abundance Factor	Zybailov et al. (2006)	○	X	N/A
PLGEM	Power Law Global Error Model	Pavelka et al. (2008)	○	X	N/A
SI_N	Normalized Special Index	Griffin et al. (2010)	○	X	N/A
SpI	Special Index	Fu et al. (2008)	○	○	Y
Qspec	Qspec	Choi et al. (2008)	○	○	Y
Hybrid	Hybrid approach	Booth et al. (2011)	○	○	Y

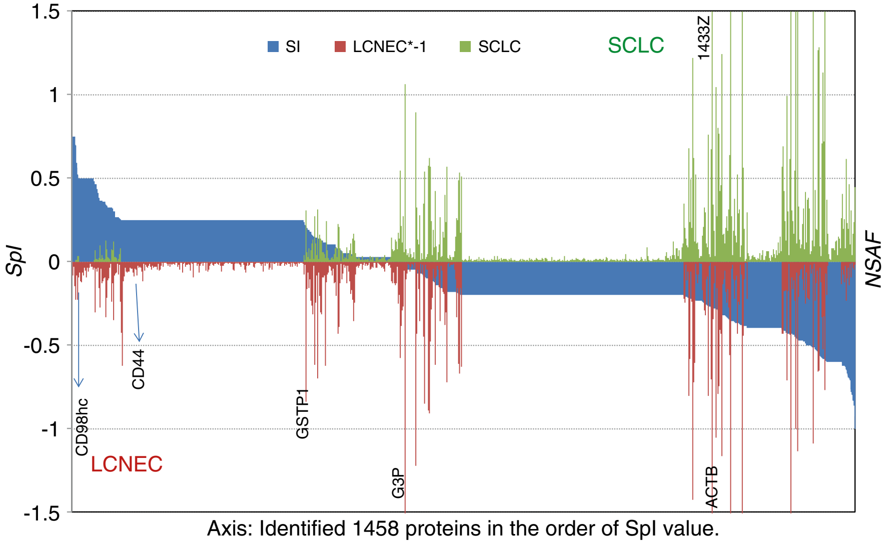


Fig. 5.11 The *SpI*-based comparison between the large-cell neuroendocrine lung cancer (LCNEC) (n=4) and the small-cell lung cancer (SCLC) (n=5). Proteins were identified from their FFPE tissue specimens

and/or sample sizes. A proteomic comparison is exemplified in Fig. 5.11, where the *SpI*-based method is applied to proteins identified in the FFPE tissue specimens surgical resected from the patients of the two cancer phenotypes, the large-cell neuroendocrine lung carcinomas (LCNEC) (n=4) and small-cell lung cancers (SCLC) (n=5) (data from Nomura et al. (2011)).

5.4 Bioinformatics for Sequence Identification

Protein identification in shotgun proteomic approaches (bottom-up) has several issues. One comes from noises in intensity (y-axis) of the observed tandem mass spectra, whereas the *m/z* values (x-axis) are considerably accurate in recent high-resolution mass spectrometers. Fragment ion mass spectra acquired often miss some sequence ions and contain a variety of unexplained ion peaks, which might come from unusual fragmentation and fragmentation of co-eluting peptides having the m/z values within an isolation mass window of precursor ions to be sequenced. There are now four peptide sequencing strategies using MS/MS spectra: database search, Spectral library matching, *de novo* sequencing, hybrid approaches using sequence-tag determination followed by database search, as illustrated in Fig. 5.12.

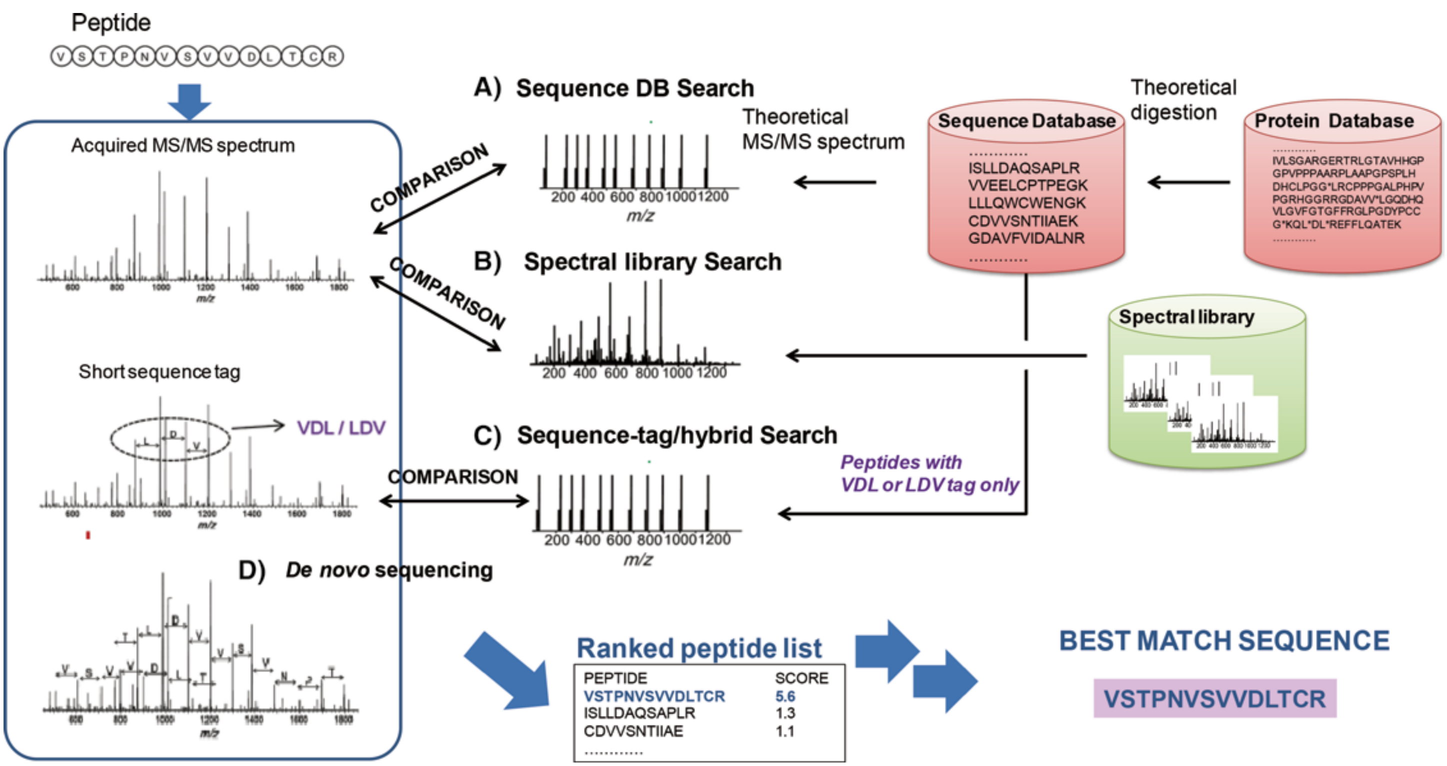


Fig. 5.12 An overview of peptide identification. Peptides can be identified by the following methods: matching acquired MS/MS spectra with theoretical spectra of peptides generated form a protein database (*A*, database search), or with spectra deposited in a spectral library (*B*, spectral library search); (*C*) hybrid methods such as sequence tag-assisted database search, which uses short sequence tags (3 amino acids in this illustration) extracted from real spectra to pick up peptides with those tag by *A*; (*D*) *de novo* sequencing of acquired spectra (Adapted with modification from Nesvizhskii 2010)

5.4.1 Database Search Approach

Database search methods are most commonly adapted by proteomic research laboratories. In principle, when a tandem mass spectrum (S^{obs}) is obtained for a precursor peptide mass (m^{obs}) with a specified tolerance (δ^{spec}), a database search algorithm extracts all peptides within the mass tolerance range from the database (P^{spec}):

$$\{p\}_{m^{obs}} = \left\{ p : p \in P^{spec};\ \left| m^{p} - m^{obs} \right| < \delta \right\}, \tag{5.3}$$

where p is a candidate peptide and m^p its theoretical mass. $\{p\}_{m^{obs}}$ is the ensemble of candidate peptides, and usually comprises of thousands of candidate peptides whereas the ensemble's size depends on the database size (P^{spec}) and δ^{spec}. Then, a theoretical fragmentation mass spectrum (S_p^{calc}) of each candidate peptide (p) is compared to the observed spectrum (S^{obs}) using a score function, $\Gamma(S^{obs}, S_p^{calc}; m^{obs}, m^p;\ \cdots\cdots)$. Database search algorithms utilize their unique score functions to evaluate a match of S^{obs} and S_p^{calc}, and reports the list of candidate peptides in the order of better scores. The score function attempts to calculate the degree of similarity between the observed MS/MS spectrum and the theoretical spectrum. There are various search algorisms with each dedicated scoring method such as SEQUEST (Eng et al. 1994), X! Tandem (Craig and Beavis 2004), OMSSA (Geer et al. 2004), MASCOT (Perkins et al. 1999), SpectrumMill (www.chem.agilent.com/), and PHENYX (Colinge et al. 2003). The reported scores, e.g., cross-correlation score (*XCorr)* of SEQUEST can be converted to a general statistical measure such as *p*-value or the expectation value, *E*-value (Fenyo and Beavis 2003).

5.4.2 Spectral-Library Matching Approach (Tharakan et al. 2010; Ning et al. 2010; Ahrne et al. 2009)

Library search methods use a spectral library compiled from a large collection of peptide MS/MS spectra that are previously observed and identified experimentally. An unknown spectrum is compared with all candidate spectra in the library for the best match under appropriate statistical criteria. This method has been the gold standard for MS analysis of small molecules, but its application to proteomics has become possible very recently because of reliable construction of peptide MS/MS spectra with an increasing proteomic data deposition in public library. As long as such a library is correctly constructed from experimental MS/MS spectra with their high purities, searching an observed MS/MS spectrum against the library gives a confident identification compared with theoretical spectra.

5.4.3 De novo *Spectrum Sequencing (Seidler et al. 2010; Frank et al. 2007; Pan et al. 2010; Frank and Pevzner 2005; Ma et al. 2003; Tanner et al. 2005; Dasari et al. 2010)*

The *de novo* sequencing approach is quite useful when a sequence is not present in databases and spectral libraries. Currently there are several tools available but *de novo* sequencing analysis is not yet practically applicable for large-scale data analysis such as disease-related clinical proteomic studies. The *de novo* sequencing requires a high computational capability dealing with a great number of mathematical amino-acid combinations, and high quality MS/MS spectra should be acquired. Recent *de novo* sequencing methodologies are developed on highly accurate MS/MS spectra acquired using high mass accuracy instruments under fragmentation methods such as HCD, ETD and/or a combination of CID and ETD (Ahrne et al. 2009; Frank et al. 2007).

5.4.4 *Sequence-Tag/Hybrid Approaches*

A combination of *de novo* sequencing with database searching has been developed. An internal short part of the full peptide sequence, "sequence tag" can be determined by de novo sequencing of an acquired MS/MS spectrum. Such a sequence-tag has a prefix mass and a suffix mass values which designate its position within the peptide sequence (Tanner et al. 2005). Then, each observed MS/MS spectrum is subjected to database search only against those candidate peptides containing the targeted sequence tags, which reduces the search time by restricting the number of comparisons. Alternatively, the longer subsequences that are extracted using de novo methods, called "spectral dictionary" or "gapped peptides", (Dasari et al. 2010) can be used to search against the sequence database. Such hybrid approaches would be useful for the identification of post-translationally or chemically modified peptides (Kim et al. 2009).

5.4.5 *Statistical Confidence on Peptide Identification*

Numerous matches are individually created for observed MS/MS spectra by the database search methods but only parts of peptide spectrum matches (PSMs) are correct. Therefore, assessing the confidence of PSMs and estimating the error rates on filtering a PSM list is quite crucial. The *p*-values or *E*-values (Kim et al. 2008) computed from the original scores are single-spectrum statistical confidence measures. Regardless of scoring methods, those values are fairly invariant and provide an evident validity of the identifications across different instruments and search algorithms. For the analysis involving simultaneous processing of multiple MS/MS spectra against a large protein database, we will get some of the p-values obtained to be small only by chance. Hence, a multiple testing correction of the individually computed *p*-values is needed to account for a large collection of PSMs. In order to achieve a specified overall low error rate, the threshold *p*-value is adjusted using classical adjustments such as "Bonferroni correction", (Dudoit et al. 2002) which is a family-wise error rate measure but is quite conservative for large size of datasets. Therefore, another kind of statistical measures is required for filtering of very large collections of PSMs.

5.4.5.1 False Discovery Rate (FDR)

The false discovery rate (FDR) has been introduced by Benjamini and Hochberg (Benjamini and Hochberg 1995) as the well accepted statistical confidence measure for the entire collection of PSMs, and is not associated with the individual PSM. A 1 % FDR threshold means that 99 % matches are correct and 1 % is incorrect in a PMS list accepted using a search-engine score threshold. The computation of FDR is based on the analysis of global distribution of PSM scores in the entire multiple MS/MS spectra. We can use the target-decoy database search strategy to compute FDR in MS/MS-based proteomics (Elias and Gygi 2007). In this strategy experimental MS/MS spectra are searched against a target database of protein sequences appended with the reversed (randomized, or shuffled) sequences of the same size, or separately searched against the target and decoy sequences (Lam et al. 2010). The basic assumption is that both matches to decoy peptide sequences (decoy PSMs) and false matches to sequences from the target database follow the same distribution. Then, PSMs are filtered using various score cut-offs depending on the search engines, and the corresponding FDR for each cut-off is estimated as N_d/N_t or $2N_d/(N_t+N_d)$. Herein, N_t is the number of target PSMs with scores above the cut-off, and N_d is the number of decoy PSMs among them. It is assumed that the number of incorrect target PSMs ($N_{incorrect}$) can be estimated as the number of decoy PSMs when the equal size of the target and decoy database is given. In contrast to FDR, the q-value is a confidence measure of the individual PMS for a peptide in a PMS list. A q-value of 0.01 for a given peptide matching to a target spectrum means that the 1 % FDR threshold is at least required for this PMS to be present in the PMS list. The FDR calculation using statistical models have also been reported, which does not need the decoy search.

5.4.5.2 Machine Learning Methods

Machine learning methods have been applied as post-processors that discriminate between correct and incorrect identifications. A utilization of methods such as support vector machines (SVM) (Anderson et al. 2003), linear discriminant analysis

(Keller et al. 2002), or decision trees (Elias et al. 2004), have demonstrated the considerable increase in the number of spectra confidently identified from a given shotgun proteomic experiment. Furthermore, semi-supervised learning methods (Käll et al. 2007; Choi and Nesvizhskii 2008) have shown a great improvement of the ranking process for peptides identified in complex MS/MS spectra on the basis of the characteristics of a given data set. Percolator is one of machine learning approaches (Käll et al. 2007) and a post-search engine processor that can now use output results of Sequest (Käll et al. 2007), Mascot (Brosch et al. 2009) and X!Tandem. Percolator reduces the dependence on the training data via a dynamic learning approach (Käll et al. 2007). The use of dynamic training and direct optimization also has been recently developed by other groups with a specific focus on phosphorylated peptides (Du et al. 2008). Finally, improved discrimination can be achieved by combining the output from two or more different database.

5.4.6 Protein Inference Problem (Nesvizhskii and Aebersold 2005; Nesvizhskii 2010; Li and Radivojac 2012)

Shotgun proteomic approaches are powerful in capturing proteins/peptides expressed in complex samples, currently and even in future. Recent mass spectrometric developments have allowed great speeding up of data acquisition and made possible to generate a large number of datasets collected in multiple replicates. Clinical proteomic studies are concerned with generation of highly overlapping datasets across multiple samples. However, biomedical conclusions resulting from such a large number of protein identifications would be negatively impacted by the presence of mis-identified proteins. A quite important issue regarding to various aspects of proteomic data analysis is filtering and assessing error rates at the protein levels in large datasets obtained by shotgun proteomics.

To attain a correct list of identified proteins from a set of peptide sequences with their identification scores does seem not to be straightforward, and remains quite challenging due to the following major difficulties: (1) loss of precise connectivity between peptides obtained by protein digestion and original proteins raises the protein inference problem, (Nesvizhskii and Aebersold 2005; Nesvizhskii 2010; Li and Radivojac 2012; Nesvizhskii et al. 2003) which is the very intrinsic problem for shotgun proteomics. In other words, many peptide sequences can be originated from more than one protein in a database (*e.g. degenerate or shared peptides*); (2) Identification of peptides is not reproducible actually in proteomic experiments, which probability is then referred to as "*peptide detectability*". (Tang et al. 2006); (3) a large number of low-scoring PSMs are troublesome in determining the reliable identification of peptides/proteins; most importantly, (4) how to compute protein-level FDRs still remains open and not trivial issues even though peptide-level FDRs can be obtained by a well-characterized method (*e.g. non-trivial estimation of FDRs*); (5) the identified peptide coverage for each protein are still low; (6) a growing propagation of error rates of PMS by grouping of peptides at protein levels. Various computational approaches attacking to these "*protein inference problem*" have been proposed, but the details of these methods are theoretically complex (Baldwin 2004; Zhang et al. 2007; Ma et al. 2009; Wilkins et al. 2006) and those are rather not helpful for clinical proteomics practitioners. We here just introduce some software useful for practical analysis.

1. Methods based on rules that rely on a relatively small set of confidently identified (unique) peptides that are subsequently assigned to proteins (Baldwin 2004).
2. Methods using combinatorial optimization algorithms by which the protein inference problem is solved under constrained optimization formulations and minimal protein lists covering some or all confidently identified peptides are obtained (Zhang et al. 2007; Ma et al. 2009).
3. Methods based on probabilistic inference algorithms that assign identification probabilities for each protein in a database.

The first rule-based approach is relative simple but obviously is limited in their performance since there is no rigorous criterion for the combination of the peptide identification scores and prior knowledge. The second approaches using combinatorial optimization algorithms typically take inputs of both a set of confidently identified peptides and a protein database. To obtain minimal protein lists, it therein is to solve the problem so that some or all peptides confidently identified can be covered under optimizing a certain criteria. Such formulations, called minimum set cover (MSC), are the Non-deterministic Polynomial-time hard (NP-hard) problems, a class of problems in computational complexity theory (Zhang et al. 2007; Ma et al. 2009; Wilkins et al. 2006). The issue remains for its applicability for co-expression of protein isoforms in biological samples, and IDPicker (Zhang et al. 2007; Ma et al. 2009) ignores other information including unidentified peptides, and proteins identified by shared or degenerate peptides are *indistinguishable*. The third probabilistic approaches generally consist of two steps. PSM scores are firstly converted to PSM probabilities using algorithms, and then protein inference is performed based on an assumed probabilistic model. Several probabilistic algorithms proposed include ProteinProphet, (Nesvizhskii et al. 2003) MSBayesPro, (Li et al. 2009) Fido, (Serang and Noble 2012; Serang et al. 2010) and MIPGEM, (Shen et al. 2008) which have different strategies and levels of rigor in addressing protein groups and different run-time performance. Algorithms are different in dealing with degenerate peptides and protein grouping for indistinguishable proteins.

The protein inference problem, protein identification, still remains to be developed. Furthermore, it has been suggested that a better quantitative estimation of peptide/protein might also help protein inference by improving the quantity adjustment of peptide detectability (Li and Radivojac 2012; Li et al. 2009), and provide additional input information for protein inference. Protein inference can be viewed as a special case of protein label-free quantification. An ideal inference algorithm should be tightly connected to a quantification algorithm. It is believed that much better performance can be achieved by combining the protein inference and quantification algorithms into one statistical framework. A fast and rigorous probabilistic inference algorithm with controllable error bound is needed.

5.5 Perspectives

Technological advance in mass spectrometry enables high-throughput shotgun proteomic experiments to produce thousands of tandem-MS spectra in exploratory clinical studies as illustrated in this chapter (Wolters et al. 2001; Hortin et al. 2010; Anderson 2010b). However, there is no proteomics technology platform that handles the many millions of human proteins expressing in tissues during a set of experiments, and researchers still have to improve shotgun proteomic technology so as to definitely archive identification and quantification of clinically important low-abundance proteins. One of the most serious problems of the present shotgun proteomics is poor ability to identify low-abundance proteins, even though the instrument's sensitivity is enough to detect their precursor ions. This is because MS/MS spectra cannot be acquired or MS/MS spectra are indeed acquired, but those are only partially used for identification of peptides and proteins.

In shotgun proteomics, data-dependent acquisition (DDA) of MS/MS spectra is currently popular, but it is prone to identify abundant peptides and thus limit the detectable dynamic range. Michalski A., et al. (2011) have pointed out that only about 16 % of detectable peptides had been targeted by DDA although more than 100,000 detectable peptides elute in a single shotgun LC-MS/MS runs, indicating that the majority of peptide ion signals remains inaccessible.

To overcome these problems, data-independent acquisition (DIA) methods have recently been developed, and will open a new landscape on peptide/protein identification and quantification in exploratory experiments (Panchaud et al. 2011; Geiger et al. 2010; Gillet et al. 2012). A typical data-independent scan sequence consists of the isolation and subsequent fragmentation of successive windows of about 10 m/z throughout the desired mass range. Such an innovative DIA-based shotgun system needs a high speed scanning cycle and optimal instrumentation with a high MS resolution, and also need to develop efficient software to deconvolute mixture MS/MS spectra generated from a wider isolation windows. New software along this line has recently started to be developed (Bern et al. 2010; http://www.physikron.com/technology/; Egertson et al. 2013). In the near future implementation of DIA technology into clinical biomarker studies will become greatly powerful in mining biomarker candidates.

References

Abderhalden E, Brockmann H. The contribution determining the composition of proteins especially polypeptides (German). Biochem Z. 1930;225:386–408.

Ahrne E, Masselot A, Binz PA, Muller M, Lisacek F. A simple workflow to increase MS2 identification rate by subsequent spectral library search. Proteomics. 2009;9:1731–6.

Anderson NG. Adventures in clinical chemistry and proteomics: a personal account. Clin Chem. 2010a;56:154–60.

Anderson NL. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clin Chem. 2010b;56:177–85.

Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1:845–67.

Anderson DC, Li W, Payan DG, Noble WS. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and sequest scores. J Proteome Res. 2003;2:137–46.

Baldwin MA. Protein identification by mass spectrometry: issues to be considered. Mol Cell Proteomics. 2004;3(1):1–9.

Benjamini Y, Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J R Stat Soc Ser B-Methodol. 1995;57:289–300.

Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ, MacCoss MJ. Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. Anal Chem. 2010;82:833–41.

Biemann K. Contributions of mass spectrometry to peptide and protein structure. Biomed Environ Mass Spectrom. 1988;16:99–111.

Biemann K, Papayannopoulos IA. Amino acid sequencing of proteins. Acc Chem Res. 1994;27:370–8.

Booth JG, Eilertson KE, Paul DB, Olinares HY. A Bayesian mixture model for comparative spectral count data in shotgun proteomics. Mol Cell Proteomics. 2011;10:M110.007203.

Brosch M, Yu L, Hubbard T, Choudhary J. Accurate and sensitive peptide identification with mascot percolator. J Proteome Res. 2009;8:3176–81.

Choi H, Nesvizhskii AI. Semi-supervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res. 2008;7:254–65.

Choi H, Fermin D, Nesvizhskii AI. Significance analysis of spectral count data in labelfree shotgun proteomics. Mol Cell Proteomics. 2008;7:2373–85.

Chung TW, Tureček F. Backbone and side-chain specific dissociations of z ions from non-tryptic peptides. J Am Soc Mass Spectrom. 2010;21:1279–95.

Chung TW, Tureček F. Proper and improper aminoketyl radicals in electron-based peptide dissociations. Int J Mass Spectrom. 2011;301:55–61.

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics. 2003;3:1454–63.

Coon JJ. Collisions or electrons? Protein sequence analysis in the 21st century. Anal Chem. 2009;81:3208–15.

Coon JJ, Syka JEP, Schwartz JC, Shabanowitz J, Hunt DF. Anion dependence in the partitioning between proton and electron transfer in ion/ion reactions. Int J Mass Spectrom. 2004;236:33–42.

Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20:1466–7.

Csonka IP, Paizs B, Lendvay G, Suhai S. Proton mobility in protonated peptides: a joint molecular orbital and RRKM study. Rapid Commun Mass Spectrom. 2000;14:417–31.

Csonka IP, Paizs B, Lendvay G, Suhai S. Proton mobility and main fragmentation pathways of protonated lysylglycine. Rapid Commun Mass Spectrom. 2001;15:1457–72.

Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJL, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res. 2010;9:1716–26.

Du X, Yang F, Manes NP, Stenoien DL, Monroe ME, Adkins JN, et al. Linear discriminant analysis-based estimation of the false discovery rate for phosphopeptide identifications. J Proteome Res. 2008;7:2195–203.

Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sin. 2002;12:111–39.

Edman P. A method for the determination of the amino acid sequence in peptides. Arch Biochem. 1949;22:475–6.

Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, Canterbury JD, Marsh DM, Kellmann M, Zabrouskov V, Wu CC, MacCoss MJ. Multiplexed MS/MS for improved data-independent acquisition. Nat Methods. 2013;10(8):744–6.

Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007;4:207–14.

Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol. 2004;22:214–9.

Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J Am Soc Mass Spectrom. 1994;5:976–89.

Engel BJ, Pan P, Reid GE, Wells M, McLuckey SA. Charge state dependent fragmentation of gaseous protein ions in a quadrupole ion trap: bovine ferri-, ferro-, and apo-cytochrome c. Int J Mass Spectrom. 2002;219:171–87.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989;246(4926):64–71.

Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry based protein identifications using general scoring schemes. Anal Chem. 2003;75:768–74.

Frank A, Pevzner P. PepNovo: De novo peptide sequencing via probabilistic network modeling. Anal Chem. 2005;77:964–73.

Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA. De novo peptide sequencing and identification with precision mass spectrometry. J Proteome Res. 2007;6:114–23.

Frese CF, Maarten Altelaar AF, van den Toorn H, Nolting D, Griep-Raming J, Heck AJR, Mohammed S. Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Anal Chem. 2012;84:9668–73.

Fu X, Gharib SA, Green PS, Aitken ML, Frazer DA, Park DR, et al. Spectral index for assessment of differential protein expression in shotgun proteomics. J Proteome Res. 2008;7:845–54.

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open mass spectrometry search algorithm. J Proteome Res. 2004;3:958–64.

Geiger T, Cox J, Mann M. Proteomics on an Orbitrap Benchtop mass spectrometer using all-ion fragmentation. Mol Cell Proteomics. 2010;9:2252–61.

Gilbert RG, Smith SC. Theory of unimolecular and recombination reactions. Oxford: Blackwell Scientific Publications; 1990. p. 52–132.

Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012;11:1–17.

Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotech. 2010;28:83–9.

Gu C, Tsaprailis G, Breci L, Wysocki VH. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. Anal Chem. 2000;72:5804–13.

Gundry RLFQ, Jelinek CA, Van Eyk JE, Cotter RJ. Investigation of an albumin-enriched fraction of human serum and its albuminone. Proteomics Clin Appl. 2007;1:73–88.

Harrison AG. Linear free energy correlations in mass spectrometry. J Mass Spectrom. 1999;34:577–89.

Harrison AG, Yalcin T. Proton mobility in protonated amino acids and peptides. Int J Mass Spectrom Ion Process. 1997;165:339–47.

Hill EG, Schwacke JH, Comte-Walters S, Slate EH, Oberg AL, Eckel-Passow JE, Terry M, Therneau TM, Schey KL. A statistical model for iTRAQ data analysis. J Proteome Res. 2008;7:3091–101.

Hood BL, Darfer MM, Furusato B, Lucas DA, Ringeisen BR, Sesterhenn IA, et al. Proteomic analysis of formalin-fixed prostate cancer tissue. Mol Cell Proteomics. 2005;4:1741–53.

Hortin GL, Carr SA, Anderson NL. Introduction: advances in protein analysis for the clinical laboratory. Clin Chem. 2010;56:149–51.

Käll L, Canterbury J, Weston J, Noble WS, MacCoss MJ. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. Nat Methods. 2007;4:923–5.

Karas M, Hillenkamp F. Laser desorption ionization of protein with molecular masses exceeding 10,000 daltons. Anal Chem. 1988;60:2299–301.

Kawamura T, Nomura M, Tojo H, Fujii K, Hamasaki H, Mikami S, Bando Y, Kato H, Nishimura T. Proteomic analysis of laser-microdissected paraffin-embedded tissues: (1) stage-related protein candidates upon non-metastatic lung adenocarcinoma. J Proteomics. 2010;73:1100–10.

Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. Anal Chem. 2002;74:5383–92.

Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J Proteome Res. 2008;7:3354–63.

Kim S, Bandeira N, Pevzner PA. Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. Mol Cell Proteomics. 2009;8:1391–400.

Kovács A, Sperling E, Lázár J, Balogh A, Kádas J, Szekrényes Á, Takács L, Kurucz I, Guttman A. Fractionation of the human plasma proteome for monoclonal antibody proteomics-based biomarker discovery. Electrophoresis. 2011;32:1916–25.

Kruger NA, Zubarev RA, Horn DM, McLafferty FW. Electron capture dissociation of multiply charged peptide cations. Int J Mass Spectrom. 1999;187:787–93.

LaBaer J. Improving international research with clinical specimens: 5 achievable objectives. J Proteome Res. 2012;11:5592–601.

Lam H, Deutsch EW, Aebersold R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. J Proteome Res. 2010;9:605–10.

Li YF, Radivojac P. Computational approaches to protein inference in shotgun proteomics. BMC Bioinform. 2012;13 Suppl 16:S4.

Li YF, Arnold RJ, Li Y, Radivojac P, Sheng Q, Tang H. A Bayesian approach to protein inference problem in shotgun proteomics. J Comput Biol. 2009;16:1183–93.

Ma B, Zhang KZ, Hendrie C, Liang CZ, Li M, Doherty-Kirby A, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 2003;17:2337–42.

Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, et al. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. J Proteome Res. 2009;8:3872–81.

Malm J, Végvári A, Rezei M, Upton P, Danmyr P, Nilsson R, Steinfelder E, Marko-Varga G. Large scale biobanking of blood – the importance of high density sample processing procedures. J Proteomics. 2012;76:116–24.

Marcotte EM. How do shotgun proteomics algorithms identify proteins? Nat Biotechnol. 2007;25:755–7.

Marko-Varga G. BioBanking – the Holy Grail of novel drug and diagnostic developments. J Clin Bioinform. 2011;1:14.

Marko-Varga G, et al. Personalized medicine and proteomics: lessons from non-small cell lung cancer. J Proteome Res. 2007;6:2925–35.

Marko-Varga G, Végvári A, Welinder C, Rezei M, Edula G, Svensson K, Belting M, Laurell T, Fehniger TE. Clinical protein science: utilization of biobank resources and examples of current applications. J Proteome Res. 2011;11:5124–34.

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, et al. PRIDE: the proteomics identifications database. Proteomics. 2005;5:3537–45.

McLafferty FW. Tandem mass spectrometry. New York: Wiley; 1983.

Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res. 2011;10:1785–93.

Morgan DG, Bursey MM. A linear free-energy correlation in the low energy tandem mass spectra of protonated tripeptides Gly–Gly-Xxx.Org. Mass Spectrom. 1994;29:354–9.

Nature. Proteomics, transcriptomics: what's in a name? Nature. 1999;402:715.

Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010;73:2092–123.

Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data – the protein inference problem. Mol Cell Proteomics. 2005;4:1419–40.

Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75:4646–58.

Ning K, Fermin D, Nesvizhskii AI. Computational analysis of unassigned high quality MS/MS spectra in proteomic datasets. Proteomics. 2010;10:2712–8.

Nomura M, Fukuda T, Fujii K, Kawamura T, Tojo H, Kihara M, Bando Y, Gazdar AF, Tsuboi M, Oshiro H, Nagao T, Ohira T, Ikeda N, Gotoh N, Kato H, Marko-Varga G, Nishimura T. Preferential expression of potential markers for cancer stem cells in large cell neuroendocrine carcinoma of the lung. An FFPE proteomic study. J Clin Bioinformatics. 2011;1:23.

O'Hair RA, Reid GE. Neighboring group versus cis-elimination mechanisms for side chain loss from protonated methionine, methionine sulfoxide, and their peptides. Eur Mass Spectrom. 1999;5:325–34.

Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, Cooper LT, Onuma OK, Spiro C, Therneau TM, Bergen III HR. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. J Proteome Res. 2008;7:225–33.

Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics. 2005;5:3226–45.

Paizs B, Suhai S. Theoretical study of the main fragmentation pathways for protonated glycylglycine. Rapid Commun Mass Spectrom. 2001a;15:651–63.

Paizs B, Suhai S. Combined quantum chemical and RRKM modeling of the main fragmentation pathways of protonated GGG. I. Cis–trans isomerization around protonated amide bonds. Rapid Commun Mass Spectrom. 2001b;15:2307–23.

Paizs B, Suhai S. Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. Rapid Commun Mass Spectrom. 2002;16:1699–702.

Paizs B, Suhai S. Towards understanding the tandem mass spectra of protonated oligopeptides. 1: mechanism of amide bond cleavage. J Am Soc Mass Spectrom. 2004;15:103–12.

Paizs B, Suhai S. Fragmentation pathways of protonated peptides. Mass Spectrom Rev. 2005;24:508–48.

Pan C, Park BH, McDonald WH, Carey PA, Banfield JF, VerBerkmoes NC, et al. A highthroughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. Bmc Bioinform. 2010;11:18.

Panchaud A, Jung S, Shaffer SA, Aitchison JD, Goodlett DR. Faster, quantitative, and accurate precursor acquisition independent from ion count. Anal Chem. 2011;83:2250–7.

Pavelka N, Fournier ML, Swanson SK, Pelizzola M, Ricciardi-Castagnoli P, Florens L, et al. Statistical similarities between transcriptomics and quantitative shotgun proteomics data. Mol Cell Proteomics. 2008;7:631–44.

Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551–67.

Polce MJ, Ren D, Wesdemiotis C. Dissociation of the peptide bond in protonated peptides. J Mass Spectrom. 2000;35:1391–8.

Prieto DA, Hood BL, Darfler MM, Guiel TG, Lucas DA, Conrads TP, et al. Liquid Tissue™: proteomic profiling of formalin-fixed tissues. Biotechniques. 2005;38:S32–5.

Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom. 1984;11:601.

Schwartz BL, Bursey MM. Some proline substituent effect in the tandem mass spectrum of protonated pentaalainine. Biol Mass Spectrom. 1992;21:92–6.

Seidler J, Zinn N, Boehm ME, Lehmann WD. De novo sequencing of peptides by MS/MS. Proteomics. 2010;10:634–49.

Serang O, Noble WS. Faster mass spectrometry-based protein inference: junction trees are more efficient than sampling and marginalization by enumeration. IEEE/ACM Trans Comput Biol Bioinform. 2012;2012.

Serang O, MacCoss MJ, Noble WS. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J Proteome Res. 2010;9:5346–57.

Shen C, Wang Z, Shankar G, Zhang X, Li L. A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. Bioinformatics. 2008;24:202–8.

Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol. 2004;5:699–711.

Summerfield SG, Cox KA, Gaskell SJ. The promotion of d-type ions during the low-energy collision-induced dissociation of some cysteic acid-containing peptides. J Am Soc Mass Spectrom. 1997;8:25–31.

Swaney DL, McAlister GC, Wirtala M, Schwartz JC, Syka JE, Coon JJ. Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. Anal Chem. 2007;79:477–85.

Swaney DL, McAlister GC, Coon JJ. Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat Methods. 2008;5:959–64.

Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A. 2004;101:9528–33.

Syrstad EA, Tureček F. Hydrogen atom adducts to the amide bond. Generation and energetics of the amino(hydroxy)methyl radical in the gas phase. J Phys Chem. 2001;A105:11144–55.

Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T. Protein and polymer analyses up to 100,000 by laser ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom. 1988;2:151–3.

Tang X, Thibault P, Boyd RK. Fragmentation reactions of multiplyprotonated peptides and implications for sequencing by tandem mass spectrometry with low-energy collision-induced dissociation. Anal Chem. 1993;65:2824–34.

Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, Novotny MV, Reilly JP, Radivojac P. A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics. 2006;22:e481–8.

Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, et al. InsPecT: identification of post-translationally modified peptides from tandem mass spectra. Anal Chem. 2005;77:4626–39.

Tharakan R, Edwards N, Graham DRM. Data maximization by multipass analysis of protein mass spectra. Proteomics. 2010;10:1160–71.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17:520–5.

Tsaprailis G, Nair H, Somogyi Á, Wysocki VH, Zhong W, Futrell JH, Summerfield SG, Gaskell SJ. Influence of secondary structure on the fragmentation of protonated peptides. J Am Chem Soc. 1999;121:5142–54.

Tureček F, Syrstad EA. Mechanism and energetics of intramolecular hydrogen transfer in amide and peptide radicals and cation-radicals. J Am Chem Soc. 2003;125:3353–69.

Vaisar T, Urban J. Probing the proline effect in CID of protonated peptides. J Mass Spectrom. 1996;31:1185–7.

Végvári A, Rezeli M, Döme B, Fehniger TE, Marko-Varga G. Translation science for targeted personalized medicine treatments. In: Sanders S, editor. Selected presentations from the 2011 Sino-american symposium on clinical and translational medicine. Washington, DC: Science/AAAS; 2011a. p. 36–7.

Végvári Á, Welinder C, Lindberg H, Fehniger TE, Marko-Varga G. Biobank resources for future patient care: developments, principles and concepts. J Clin Bioinform. 2011b;1:24.

Wilkins MR, Appel RD, Van Eyk JE, Chung MC, Gorg A, Hecker M, Huber LA, Langen H, Link AJ, Paik YK, et al. Guidelines for the next 10 years of proteomics. Proteomics. 2006;6(1):4–8.

Wisniewski JR, Ostasiewicz P, Mann M. High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. J Proteome Res. 2011;10:3040–9.

Wolters DA, Washburn MP, Yates III JR. An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem. 2001;73:5683–90.

Wysocki VH, Tsaprailis G, Smith LL, Breci LA. Mobile and localized protons: a framework for understanding peptide dissociation. J Mass Spectrom. 2000;35:1399–406.

Yamashita M, Fenn JB. Negative ion production with the electrospray ion source. J Phys Chem. 1984;88:4671–5.

Yao C, Syrstad EA, Tureček F. Electron transfer to protonated beta-alanine N-methylamide in the gas phase: an experimental and computational study of dissociation energetics and mechanisms. J Phys Chem A. 2007;111:4167–80.

Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng. 2009;11:49–79.

YuW VJE, Huberty MC, Martin SA. Identification of the facile gasphase cleavage of the Asp-Pro and Asp-Xxx peptide bonds in matrix assisted laser desorption time-of-flight mass spectrometry. Anal Chem. 1993;65:3015–23.

Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res. 2007;6:3549–57.

Zubarev RA, Kelleher NL, McLafferty FW. Electron capture dissociation of multiply charged protein cations – a nonergodic process. J Am Chem Soc. 1998;1998(120):3265–6.

Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res. 2006;5:2339–47.

Chapter 6
Post-translational Modifications in the Human Proteome

Cheryl F. Lichti, Norelle C. Wildburger, Mark R. Emmett, Ekaterina Mostovenko, Alexander S. Shavkunov, Shinji K. Strain, and Carol L. Nilsson

Abstract The Chromosome-Centric Human Proteome Project (C-HPP) is a global project aimed to identify at least one protein isoform encoded by the approximately 20, 300 human genes. In addition, protein post-translational modifications will be characterized, with the initial goal of detecting phosphorylation, acetylation, and glycosylation sites in each protein. In this chapter, we provide an overview of known post-translational modifications, their known biological functions, and present strategies to detect them on both a single protein and proteomic scales. In future proteomic studies, global characterization of post-translation modifications, splice variants, and variants caused by single nucleotide polymorphisms (SNPs) will be necessary to fully understand the role of proteins in human biology and disease.

Keywords Post-translational modification • Phosphorylation • Acetylation • Glycosylation • Ubiquitination • Mass spectrometry • C-HPP • Proteomics

C.F. Lichti • E. Mostovenko • A.S. Shavkunov • S.K. Strain • C.L. Nilsson (✉)
Department of Pharmacology, University of Texas Medical Branch,
301 University Blvd, Galveston, TX 77555, USA
e-mail: carol.nilsson@utmb.edu

N.C. Wildburger
Department of Pharmacology, University of Texas Medical Branch,
301 University Blvd, Galveston, TX 77555, USA

Department of Neuroscience and Cell Biology, University of Texas Medical Branch,
301 University Blvd, Galveston, TX 77555, USA

M.R. Emmett
Department of Biochemistry and Molecular Biology, University of Texas Medical Branch,
301 University Blvd, Galveston, TX 77555, USA

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_6

6.1 Introduction

The major advances in characterization of the human genome, including the first published sequence (Venter et al. 2001) and the first draft of the human genome “parts list” (Consortium 2012) promise to accelerate related projects that employ genome-wide analytical strategies. The Chromosome-centric Human Proteome Project (C-HPP) is a global research consortium that is charged with identification of all human proteins, defining their tissue and cellular expression, as well as mapping of the three major protein post-translational modifications (PTMs), acetylation, phosphorylation, and glycosylation. Because modifications change the structures and functions of proteins, systems-wide characterization of protein PTMs can significantly advance our knowledge of the role of PTMs in human development, health and disease.

The type and number of PTMs that have been characterized are highly diverse structurally and number over 100. The discoveries of new modifications are still being described in the scientific literature. While the goal of protein characterization within the C-HPP is limited to the description of protein acetylation, phosphorylation and glycosylation, there are other PTMs that can be captured on a proteomic scale, including modification by ubiquitin or small ubiquitin-like modifications (SUMO), lipidation, nitration, and even halogenation.

Protein PTMs are often dynamic, changing in response to intracellular or extracellular stimuli, to developmental signals or aging, in distinct tissue localizations or globally (Fig. 6.1). Furthermore, the interplay of systems of PTMs, including

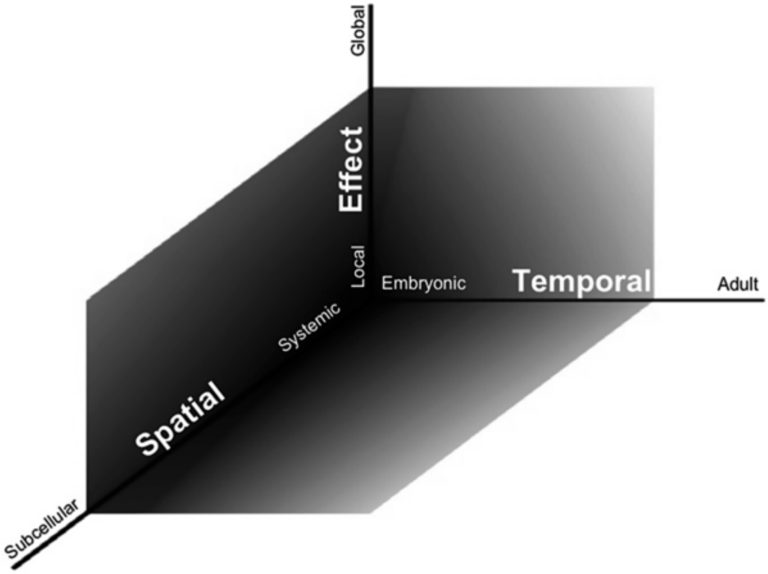


Fig. 6.1 Protein PTMs are dynamic, changing locally or globally in response to developmental signals, normal stimuli, or disease processes

Table 6.1 The protein families that control protein functions mediated by key PTMs. Examples of writers, readers and erasers as key control proteins in biological systems

Writers	Readers	Erasers
Histone acetylases	BRD proteins	Histone deacetylases
Kinases	SH_2-domain containing proteins	Phosphatases
Glycosyltransferases	Lectins	Glycosidases
Ubiquitin ligases	UB binding proteins	Deubiquitinases

The interplay of their activities contributes to the complexity of biochemical signaling pathways in health and disease states. Thus, integrated biological studies greatly benefit from defining phosphorylation, glycosylation and ubiquitin-mediated signaling

acetylation, phosphorylation, glycosylation and ubiquitination adds more layers to the complexity of signaling through those modifications. The fine regulation of protein PTMs is provided by key control proteins (Table 6.1), the so-called writers, readers and erasers of the chemical language embedded in PTMs.

Another characteristic of PTMs is that they are typically substoichiometric in proteolytic mixtures. Identification and site localization of PTMs often requires one or more enrichment steps. In this chapter, we describe functions associated with acetylation, phosphorylation, glycosylation and a few other selected PTMs. Further, analytical approaches to determine site localization and types of modifications on a proteomic scale are described.

6.2 Protein Acetylation

Proteins can be either stably N-terminally acetylated or reversibly acetylated on lysine (Lys) residues. Lys acetylation plays essential roles in cell homeostasis. A large number of studies indicate that reversible acetylation of Lys is widespread in the human proteome (Lin et al. 2014). The finding that nearly every enzyme in metabolic pathways, including glycolysis, has a plastic pattern of acetylation in response to factors such as nutritional status and disease, suggests that reversible acetylation is as important to the regulation of pathways as phosphorylation, glycosylation, and ubiquitination (Kouzarides 2000; Yuan and Marmorstein 2013).

The study of histone acetylation, induced by histone acetyltransferases (HATs), has been demonstrated to have a strong correlation to the regulation of gene activation (Eberharter and Becker 2002; Kimura et al. 2005; Yang and Seto 2007). Because the site localizations of histone PTMs are important to understand their function, mass spectrometry has played a prominent role in their analysis. Histones are smaller (~20 kDa) proteins, have heterogeneous modifications, and abundant basic amino acid residues. Because the proteolytic enzyme of choice in proteomics is trypsin, which cleaves at Lys and arginine (Arg) residues, histone digests yield an abundance of fragments with variable modifications. While site localization of PTMs in the tryptic peptide derived from a complex mixture of isoforms is feasible, it is not possible to assign PTMs reliably along the full length proteins.

Fortunately, sequencing by top-down mass spectrometry of smaller proteins is ideally suited for histone studies (Siuti and Kelleher 2007; Tipton et al. 2011). In top-down mass spectrometry, multiply charged gas phase protein ions are analyzed in a Fourier transform ion cyclotron resonance mass spectrometer (Marshall and Hendrickson 2008) and fragmented by electron capture dissociation (ECD) (Kelleher et al. 1999; Zubarev et al. 2000; Zubarev et al. 1998), infrared multiphoton dissociation (IRMPD) (Little et al. 1994), or a combination of both methods (Horn et al. 2000). The advantages of ECD sequencing of large polypeptides are that site localization of PTMs is easily achieved, even for labile modifications such as phosphorylation and O-glycosylation, and that those PTMs are assigned reliably for the entire protein isoform that is analyzed as an intact molecular ion. Furthermore, the high resolution and high mass accuracy that is characteristic of FT-ICR MS allows distinction between the nearly isobaric modifications acetylation and tri-methylation ($\Delta m = 0.0364$ Da); both of those PTMs may occur on histones, along with phosphorylation. The study of histone modifications, named epigenomics, is a very active area of scientific inquiry (Rivera and Ren 2013).

Some of the enzymes identified as HATs have also been shown to acetylate non-histone proteins, but most of the protein acetyltransferases acetylate non-histone proteins exclusively (Glozak et al. 2005; Spange et al. 2009). Even though acetylation is readily detected in mass spectrometry-based assays as a mass increase of 42.01 Da of a peptide or protein and a diagnostic immonium ion at *m/z* 126.1, the number of studies of the acetylome has lagged behind those aimed to map phosphorylation and ubiquitination sites. With the revelation that reversible Lys modification by acetylation plays a large role in cellular processes, the number of reports has increased greatly.

Reversible acetylation has been reported for transcription factors and may serve as an activating or inactivating modification (analogous to the role of phosphorylation). In the case of cellular tumor antigen p53 (TP53), polyacetylation leads to DNA binding, transcriptional activity and apoptotic functions (Luo et al. 2004; Sykes et al. 2006), whereas deacetylation represses the transcriptional activity by TP53 (Luo et al. 2000; Murphy et al. 1999). Reversible acetylation also plays an essential role in DNA replication, the physical separation of chromatids during mitosis (Heidinger-Pauli et al. 2009), and traffic of RNA through the nuclear pore complex (Bannister et al. 2000; Wang et al. 2004).

N-terminal acetylation of human proteins is a widespread (roughly 80 %) co-translational modification catalyzed by N-terminal acetyltransferases (NATs). Long viewed as a mere chemical block to protein degradation, it was recently discovered that N-acetylation of proteins can act as a signal for degradation by a ubiquitin ligase (Hwang et al. 2010). Furthermore, N-terminal acetylation can serve as a signal, an alternative to protein lipidation, for subcellular localization. Acetylation of ADP-ribosylation factor-like protein 8B (ARL8B) by NatC targets the protein to the lysosomal membrane (Starheim et al. 2009). Aberrant N-terminal acetylation of hemoglobin can cause clinical diseases related to resulting abnormal oxygen-binding capacity of the holoprotein (Starheim et al. 2009; Manning et al. 2012).

Like other modified peptides, acetylated peptides may be difficult to isolate and sequence by tandem mass spectrometry (MS/MS) when they are analyzed in the background of a complex mixture. Global studies of N-terminal acetylation have been enabled by special enrichment techniques prior to analysis by liquid chromatography (LC)-MS/MS. Gevaert et al. devised a negative enrichment scheme in tryptic digests of a proteome are reacted with 2,4,6-trinitrobenzensulfonic acid. The reactive agent couples to the N-terminus of internal tryptic peptides. Upon separation of the complex mixture by C_{18}-based LC, the trinitrophenyl-bearing internal peptides are easily separated from the N-terminally acetylated peptides (Van Damme et al. 2011). Another approach that has met with success is to perform pre-separation of complex proteolytic digests by strong cation exchange chromatography (SCX) at low pH (Van Damme et al. 2011; Dormeyer et al. 2007; Crimmins et al. 1988).

6.3 Protein Phosphorylation

O-phosphorylation of protein serine (Ser), threonine (Thr), and tyrosine (Tyr) residues, a reversible process, is a well-studied mechanism of cell signaling and its regulation. Phosphorylation mainly occurs on Ser and Thr, just a few percent of phosphorylation occurs on Tyr. Protein phosphorylation is mediated by protein kinases, a superfamily of roughly 500 proteins which occupy approximately 3 % of the human genome (Manning et al. 2002). Protein phosphorylation may either increase or decrease protein activity (acting as an on-off switch), trigger a change in protein-protein binding characteristics or subcellular localization. For instance, phosphorylation of transcription factors often causes dimerization and translocation into the cell nucleus. Signal transducer and activator of transcription 3 (STAT3), a multifunctional transcription factor is archetypical in this respect (Reich 2009). The regulation of signaling pathways by protein phosphorylation has been intensively studied, especially for phospho-Tyr-mediated pathways. Historically, the bias may largely be due to the availability of reliable phospho-Tyr antibodies; antibodies for phospho-Ser and phospho-Thr antibodies typically have poorer specificity for the intended antigen. With newer, MS based sequencing strategies, thousands of phosphorylation sites can be assigned in a single experiment (Oppermann et al. 2009). Derivation of the biological meaning of changes in Ser and Thr phosphorylation can be quite challenging because relatively little has been reported in the literature on this topic.

Protein phosphorylation is typically substoichiometric, heterogeneous with respect site localization, and transitory in nature. These characteristics make MS an important tool for characterization of both single protein phosphorylation and global phosphoproteomics, because in contrast to antibody-based methods, data is acquired at both high sensitivity and specificity (Nilsson 2011a). However, there are significant challenges even in MS-based approaches. Phosphopeptides are substoichiometric and usually require an enrichment step. They are acidic in nature and thus ionize poorly in positive ion mode, in the presence of unmodified peptides (ion suppression). Furthermore, Ser- and Thr-phosphorylation are thermolabile and may dissociate with prompt loss of the phosphate group, making peptide identification and site localization of the phosphorylation site more difficult.

Sample preparation alternatives for phosphoproteomics are import to consider at the experimental planning stage. While mainly chromatographic methods are employed in LC-tandem mass spectrometry (MS/MS) workflows, gel-based methods may also be applied (Eyrich et al. 2011; Dephoure et al. 2013; Černý et al. 2013). Anion exchange methods for peptides with acidic modifications (phosphate, sialic acid, and acetylation) are particularly widespread in the literature. In SCX, a column is packed with anionic resin. A salt gradient in the mobile phases induces elution of analytes based on increasing isoelectric point. When protein or peptide mixtures pass over the column, more basic polymers are retained longer, while more acidic compounds elute off the column in the earlier fractions (Gilar et al. 2005). Thus SCX, despite its inherent low peak capacity, may benefit experiments in which enrichment of phosphorylated, acetylated or sialylated peptides is desirable (Gilar et al. 2008).

Hydrophilic interaction chromatography (HILIC) (Alpert 1990) is a separation mode that fractionates peptides or proteins based on their polarity. The approach employs a polar resin and a partly hydrophilic mobile phase. Unlike reversed phase (RP) chromatography which retains analytes based on hydrophobicity, HILIC retention is higher for hydrophilic compounds. This is important because highly hydrophilic peptides (phosphorylated and glycosylated) may not be retained at all on RP (C_{18}-based) resins, but wash off in the flow through part of the gradient. Electrostatic repulsion hydrophilic interaction chromatography or ERLIC was introduced by Alpert (2008). ERLIC is a derivation of HILIC and allows the isocratic separation of phosphopeptides and glycopeptides from a proteome digest. In ERLIC, the column matrix and analytes share the same charge resulting in electrostatic repulsion, yet the mobile phase contains enough organic solvent to force the analytes to remain on the column through hydrophilic interactions (Alpert 2008). The carboxylic acids of the C-termini, and acidic amino acid (aspartate and glutamate) side chains become protonated (–COOH) in low pH conditions (Mysling et al. 2010). These neutral ion-pairs have much higher hydrophobicity than their charged state (Ding et al. 2007; Wimley et al. 1996).

Metal- based enrichment depends on immobilized metal cations which can bind Lewis base (electron pair donating) groups on phosphates and sialic acids (Larsen et al. 2005). Immobilized metal affinity chromatography (IMAC) is widely used and requires binding of metal ions (such as Fe^{3+}, Ti^{4+}, or Zr^{4+}) to a solid surface or particle. The metal oxide approach (MOAC) is similar in nature as it requires binding of a metal oxide (TiO_2, ZrO_2) to a matrix material. The results obtained from commercially obtained MOAC media are quite varied and testing is recommended prior to large-scale studies using this enrichment approach. For reviews on this subject, two recent papers are recommended for further reading (Gates et al. 2010; Ficarro et al. 2009). IMAC is the most widely applied technique for phosphopeptide analysis, based on the number of literature references.

6.4 Protein Glycosylation

Glycosylation is necessary for protein folding, solubility, stability, trafficking, cell-cell communication, and adhesion (Varki 1993; Imperiali and Rickert 1995). It is estimated that approximately 50 % of all proteins are glycosylated, though only approximately 10 % of known proteins have been annotated as such (Apweiler et al. 1999). Unlike the proteins they modify, glycosylation synthesis is non-template driven, which produces highly complex glycan structures with large variations in branching points, linkage, monosaccharide composition, and configuration (Dell and Morris 2001). As such, it is the only PTM that requires detailed structural characterization.

There are two main types of protein glycosylation, N-linked glycosylation where the oligosaccharide is covalently attached to an asparagine (Asn) and O-linked glycosylation in which the attachment occurs on the hydroxyl group of either serine (Ser) or threonine (Thr). Protein glycosylation is tightly regulated in a series of enzymatic steps and its contribution to protein function is variable. In some instances, the lack of N-linked glycosylation targets the nascent polypeptide for degradation, while in other cases protein folding and secretion remains mildly affected if at all. N-glycans are synthesized in the endoplasmic reticulum (ER) on a dolichol donor and transferred en bloc onto nascent proteins co-translationally (Nilsson and von Heijne 1993). The first monosaccharide, N-acetylglucosamine (GlcNAc), of the trimannosyl-chitobiose core common to all N-glycans is linked via an amide bond to the asparagine (Asn) (Fig. 6.2) within the consensus sequon Asn-X-Ser/Thr (where X is any amino acid except proline) and occasionally Asn-X-Cys (Satomi et al. 2004). After attachment of the N-glycan complex, the nascent polypeptide chain enters the calnexin/calreticulon pathway, whereby the glycan complex acts as a ligand to calnexin and calreticulon, which sequesters the nascent glycopolypeptide chain for proper protein folding. Properly folded proteins leave the ER for the Golgi where the carbohydrate moiety is modified by glycosyltransferases and glycosidases (Dell and Morris 2001). O-linked glycosylation, unlike N-linked, occurs in the Golgi rather than the ER with the exception of O-mannosylation. In contrast to N-linked glycosylation, there is no consensus sequon for O-linked glycosylation, thus any Ser or Thr residue is a potential O-glycosylation site.

The close connection between changes in glycosylation and the development of cancer is well documented (Dube and Bertozzi 2005). Glycans may be over- or underexpressed compared to normal tissues and reappearance of embryonic types of glycans may re-emerge. Further, many human disorders are related to congenital disorders of glycan synthesis or degradation (CDG). While relatively rare, they serve as a reminder of the importance of glycosylation to normal function. In alpha-mannosidosis, the molecular deficit is inactivity of the enzyme responsible for hydrolyzing mannose that is alpha-linked in N-linked glycans. The clinical symptoms are defects in the immune system, abnormalities in skeletal development, hearing impairment, and mental retardation (Malm and Nilssen 2008). CDGs can also be related to deficiencies in protein sialylation, leading to severe morphogenic

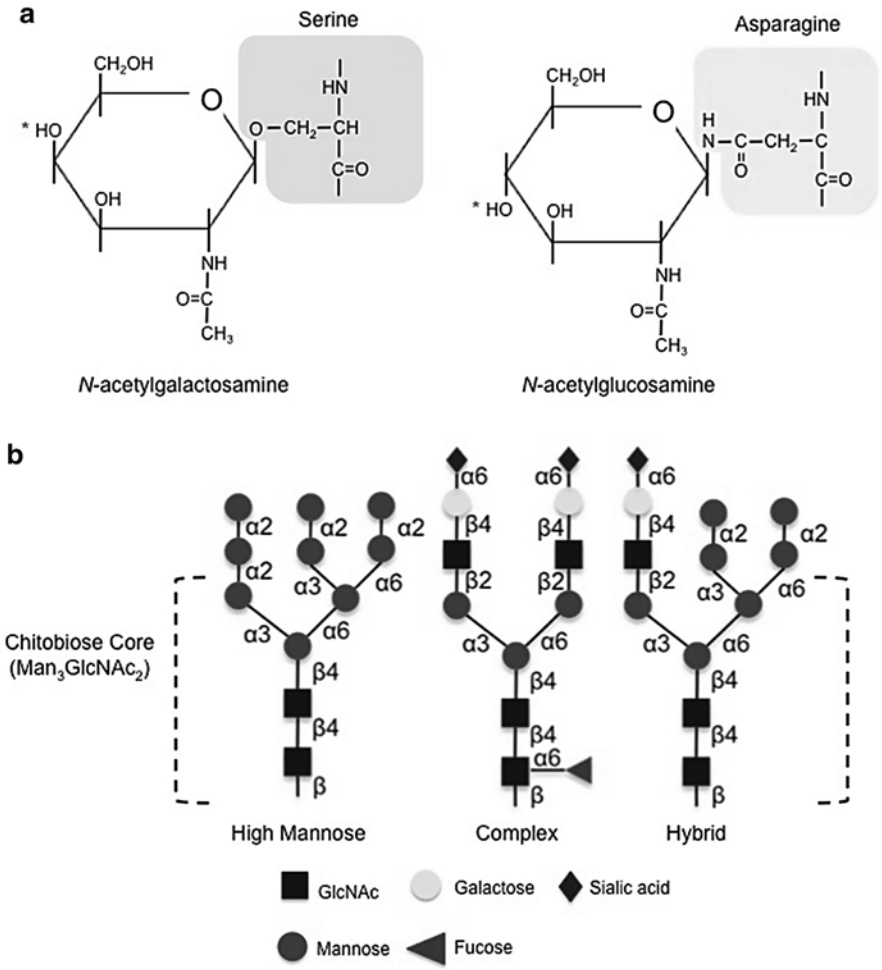


Fig. 6.2 (**a**) Structure of O- and N-linked glycan attachment to peptide (*left* and *right*, respectively). (**b**) Types of N-linked glycans

and metabolic abnormalities as well as shortened lifespan. The serum protein transferrin is normally a highly sialylated protein, lower than normal sialic acid content of this abundant protein is considered to be a biomarker for CDG (Schachter and Freeze 2009). Other defects have been linked to congenital muscular dystrophy syndromes (Martin-Rendon and Blake 2003).

Protein glycosylation is not static; its context is both spatially and temporally dependent. Though less well studied, evidence exists for the differential spatial distribution of glycoproteins. Tenascin-R, a neural extracellular matrix protein predominantly expressed in the cerebellum carries a terminal GalNAc-4-SO_4 on its

N-linked oligosaccharides (GalNAc; N-acetylgalactosamine). However within the cerebellum, only Purkinje cell bodies and their dendrites express GalNAc-4-SO_4 modified tenascin-R (Woodworth et al. 2002).

Temporally, poly(α2,8)sialic acid (PSA) is highly abundant in the brain during the early stages of development, but gradually decreases over time (Varki et al. 2009). PSA is a highly anionic homopolymer of up to 300 α2,8-linked sialic acids (Varki et al. 2009). PSA-modified neural cell-adhesion molecule at synapses reduces cell-cell and cell-matrix/extracellular adhesion likely through electrostatic repulsion by the multitude of anionic charges and large space created by the hydration volume (Varki et al. 2009), the result is a global reduction in membrane-membrane contact, effecting cell-to-cell contact with ligands, receptors, and adhesion molecules, in early development (Rutishauser 2008).

Local effects of protein glycosylation (where the precise location of the modification matters) are equally important. The receptor tyrosine kinase EPHA2 ligand Ephrin-A1 (EFNA1) contains one consensus sequon within its amino acid sequence at N26. Crystallography of the EPHA2/EFNA1 receptor-ligand complex confirmed glycosylation at N26 (Himanen et al. 2009; Himanen et al. 2010), and removal of glycosylation at this site results in the loss of EFNA1 biological function (Ferluga et al. 2013). The carbohydrate moiety of EFNA1 is essential to interact with the binding domain of EPHA2 and stabilizes the ligand-receptor interaction and subsequent tetramerization (Ferluga et al. 2013).

O-GlcNAc and phosphorylation are two different PTMs that may interact competitively, reciprocally, or simultaneously (Zeidan and Hart 2010; Hart et al. 2011). Most tumor suppressor proteins are modified by O-GlcNAc. Myc proto-oncogene protein Myc (MYC) is phosphorylated at Thr58 within the N-terminal transcriptional activation domain, which is required for MYC-dependent gene activity (Gupta et al. 1993). However, Thr58 is also modified by O-GlcNAc. Whereas phosphorylation of Thr58 promotes c-Myc transcriptional activity, the addition of O-GlcNAc likely attenuates its activity as growth inhibited cells contain predominantly O-GlcNAc modified MYC (Zeidan and Hart 2010). In another example, the addition of O-GlcNAc at Ser149 on TP53 inhibits phosphorylation at Thr155 thereby indirectly preventing ubiquitination (Yang et al. 2006). TP53 is just one example of proteins whose functions are regulated by the combination of acetylation, phosphorylation, glycosylation, and ubiquitination, underscoring the importance of mapping PTMs in the C-HPP experiments.

Glycoproteins and peptides present an analytical challenge as they are substoichiometric relative to non-glycosylated peptides, may have only partial sites of occupancy, and ionize poorly compared to their non-modified peptides (Dell and Morris 2001). These factors make it necessary to separate glycopeptides from peptides through enrichment strategies. A number of methods exist such as: lectins (Bunkenborg et al. 2004; Nilsson 2011b), hydrazide chemistry (Zhang and Aebersold 2006), graphitized carbon (Davies et al. 1992; Larsen et al. 2005), titanium dioxide (Larsen et al. 2007), strong cation exchange (Lewandrowski et al. 2007), HILIC (Hägglund et al. 2004; Wuhrer et al. 2004; Boersema et al. 2008; Di Palma et al. 2012), and ERLIC (Alpert 2007).

Lectin affinity chromatography can be used to enrich either glycoproteins or glycopeptides, but is more robust at the glycoprotein level (Atwood et al. 2006). On the other hand, highly abundant non-glycosylated proteins tend to cause 'carry over' in the glycoprotein fraction during lectin chromatography (Atwood et al. 2006; Bunkenborg et al. 2004). Lectin affinity purification at the peptide level minimizes this effect. Additional experimental considerations include the broad specificities of different lectins and the nature of the sample itself. For instance, membrane proteins require detergents or chaotropes to remain soluble, yet these reagents interfere with lectin binding; therefore, a digestion step prior to lectin enrichment is recommended.

Hydrazide chemistry captures glycoproteins on hydrazide resin after oxidation of the carbohydrate moieties (Zhang et al. 2003). Protein digestion with trypsin followed by extensive washing removes unmodified peptides. At this stage, isotopic labeling of glycopeptides for quantification occurs before removal from the hydrazide resin by PNGase F cleavage (see below). Hydrazide enrichment has limitations, including not being amenable to automation, requiring extensive reactions with toxic chemicals, prolonged enrichment, and extensive sample cleanup.

Graphitized carbon cartridges (GCC), while predominately used for the isomeric separation (Koizumi et al. 1991) of free glycans, is also applicable to the enrichment and separation of glycopeptides (Davies et al. 1992; Fan et al. 1994). GCC itself is more chemically and physically stabile than silica and can be used in a broad pH range from highly acidic to highly alkaline. The effect of temperature is not nearly as drastic as it is on silica-based resins and GCC can remove salts and detergents prior to mass spectrometry analysis (Packer et al. 1998).

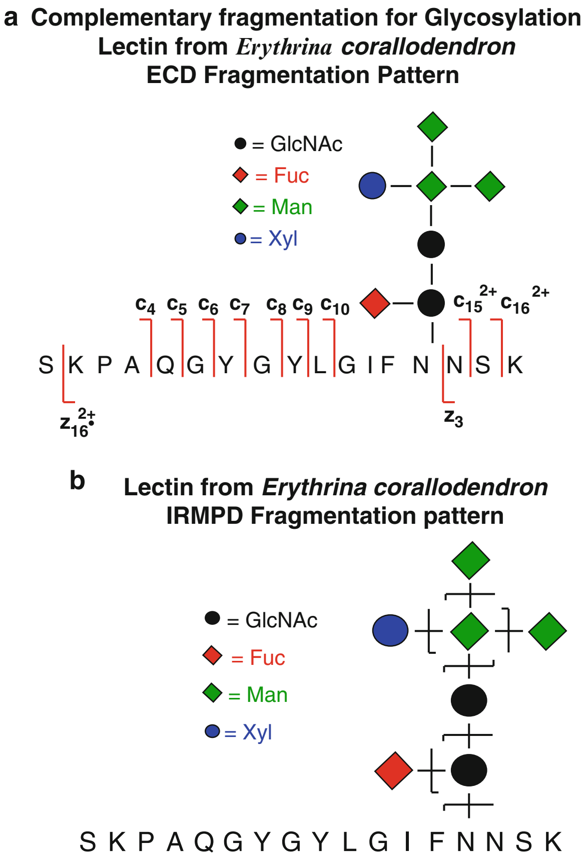
Tryptic peptides are usually too large for either enrichment (offline SPE) or separation (online nLC-MS/MS) of glycopeptides. However, GCC is well suited for smaller peptides resulting from a multienzyme or pronase digest. Pronase leads to nearly complete digestion into individual amino acids. In the case of glycopeptides, the glycan structure prevents complete enzymatic cleavage due to steric hindrance (in the case of N-linked glycopeptides) (An et al. 2003; Hua et al. 2013), resulting in a carbohydrate moiety attached to a peptide backbone up to eight amino acids long. One shortcoming of this strategy is that the peptide length may not always be optimal for database searching and protein identification.

Titanium dioxide (TiO_2) and SCX were originally applied to improve the enrichment of phosphopeptides. In recent years, Larsen and coinvestigators as well as Sickmann and colleagues adapted TiO_2 and SCX, respectively, for enrichment of terminal sialic acid containing N-linked glycopeptides (Larsen et al. 2007; Palmisano et al. 2011; Lewandrowski et al. 2007). TiO_2 (and ZrO_2)- based enrichment depends on immobilized metal cations. TiO_2 retains negatively charged sialoglycopeptides through multiple interactions with the hydroxyl and carboxyl groups (Larsen et al. 2005). However, TiO_2 also enriches phosphopeptides and peptides rich in acidic residues due to their negative charge, making it necessary to treat samples with phosphatase to prevent oversaturation of the column and improve separations. In contrast to TiO_2, sialoglycopeptides elute early in SCX fractionation. The negative charge of sialic acid counteracts the overall positive charge of the peptide (in low pH solutions) resulting in little net charge. The sialoglycopeptides carrying relatively little net charge elute in earlier fractions compared to unmodified peptides (Lewandrowski et al. 2007). SCX also enriches peptides with other acidic modifications such as phosphorylation and acetylation. Similarly to TiO_2, phosphatase treatment partially remedies this issue.

Hydrophilic interaction liquid chromatography (HILIC) is another enrichment technique to separate the glycopeptides from peptides. Glycopeptides are generally more hydrophilic than non-glycosylated peptides, which allows retention on the stationary phase through hydrophilic partitioning, based on hydrogen bonding and to some extent electrostatic interactions depending on the type of stationary phase used (Alpert 1990). However, there exists a hydrophilic overlap between non-glycopeptides and glycopeptides (Mysling et al. 2010). In complex samples, this becomes readily apparent when these peptides co-elute with glycopeptides and result in ion suppression of the glycopeptides of interest during MS analysis. The use of an ion pairing agent in the solvent system can improve the separation.

Electrostatic repulsion hydrophilic interaction chromatography or ERLIC was introduced by Alpert (Alpert 2008). The carboxylic acids of the C-termini, and acidic amino acid (aspartate and glutamate) side chains become protonated (–COOH) in low pH conditions (Mysling et al. 2010). These neutral ion-pairs have much higher hydrophobicity than their charged state (Ding et al. 2007; Wimley et al. 1996). The hydrophilic glycans (–OH groups) of the glycopeptides remain unaffected by the ion-pairing agent because of the non-ionic electrostatic interactions between the carbohydrates and stationary phase (Ding et al. 2007; Wimley et al. 1996). ERLIC, a derivation of HILIC (i.e., anion-exchange HILIC) allows the isocratic separation of phosphopeptides and glycopeptides from a tryptic proteome digest. In ERLIC, the column matrix and analytes share the same charge resulting in electrostatic repulsion, yet the mobile phase contains enough organic solvent to force the analytes to remain on the column through hydrophilic interactions (Alpert 2008). Recently several groups demonstrated the application of ERLIC for the simultaneous enrichment of glyco- and phosphopeptides in the same sample in a single chromatographic run (Zhang et al. 2010; Hao et al. 2011).

Peptide N-glycosidase F (PNGase F) hydrolyzes the β-aspartylglycosylamine bond linking the first GlcNAc of the core to the nitrogen of asparagine (N) for all N-linked glycans (except in cases of core α1,3-fucosylation). This converts the asparagine through a deamidation reaction to aspartic acid (D). Deamidation resulting from PNGase F deglycosylation yields a mass increase of +0.98 Da. This facilitates identification of N-linked glycosylation sites (when analyzed with high-resolution MS), but cannot completely guard against false positives due to spontaneous deamidation within a consensus sequence, even with isotopic (^{18}O) labeling (Robinson and Robinson 2001; Palmisano et al. 2012). This makes it absolutely necessary to have identical control samples or aliquots of the same sample without PNGase F treatment. Since glycosylation suppresses peptide ionization, it would be unlikely to see the same peptide by MS/MS if it were glycosylated. Conversely, if identified by MS/MS in the control sample, then the deamidation within the consensus sequon is most likely an artifact and a ‘false positive’. ‘True



a Complementary fragmentation for Glycosylation Lectin from *Erythrina corallodendron* ECD Fragmentation Pattern

● = GlcNAc
◆ = Fuc
◆ = Man
● = Xyl
c_4 c_5 c_6 c_7 c_8 c_9 c_{10}
c_{15}^{2+} c_{16}^{2+}
S K P A Q G Y G Y L G I F N N S K
$z_{16}^{2+\bullet}$
z_3

b Lectin from *Erythrina corallodendron* IRMPD Fragmentation pattern

● = GlcNAc
◆ = Fuc
◆ = Man
● = Xyl
S K P A Q G Y G Y L G I F N N S K

Fig. 6.3 The high degree of complementarity of ECD (**a**) and IRMPD (**b**) to derive structural data from N-linked glycopeptides

positive' sites may only be annotated as "putative' as they are observed only after enzymatic deglycosylation. An alternative strategy would be to keep the glycan moiety intact on its peptide backbone and analyze by multistage MS in an ion trap instrument (Reinhold et al. 2013), or by a combination of fragmentation techniques that fragment the glycan and the peptide backbone separately (e.g., IRMPD/ECD, CID/ETD, CID/HCD, or HCD/ETD) (Hakansson et al. 2001; Wu et al. 2007; Alley et al. 2009; Scott et al. 2011; Singh et al. 2012). Such strategies can provide complementary structural datasets (Fig. 6.3).

Alternatively, other available enzymes such as Endo H and the Endo F family (I-III) cleave within the chitobiose core between the two GlcNAc residues, leaving a single GlcNAc or fucosylated GlcNAc on the peptide. The mono- or disaccharide containing peptide can then be fragmented with MS/MS. If CID is employed, GlcNAc may be lost from the peptide backbone as an oxonium ion of 203.08 Da (or 349.14 Da in the case of fucosylated GlcNAc). Provided there is only one N-linked sequon in the peptide, site occupancy can be determined with the presence of the diagnostic oxonium ion (Hägglund et al. 2004, 2007).

In general, the same strategies described above for the analysis of N-linked glycoproteins can be applied to O-linked glycopeptides, with the exception that there is no universal enzyme, like PNGase F, for cleaving O-linked glycans. O-linked glycosylation (i.e., GalNAc) is typically shorter in length than N-linked glycosylation. Yet in many cases just as complex due to extended cores. For O-GlcNAc, due to its extreme thermolability, ETD or ECD fragmentation is recommended to determine site occupancy.

The remarkable diversity and molecular choreography of glycosylation within a spatiotemporal biological context allows cells to fine-tune the biological and biophysical properties of proteins, vastly expanding molecular communication and functional outcomes (Edwards et al. 2014). With the ever-increasing knowledge of protein glycosylation and other PTMs, the next major and necessary task will be the large-scale elucidation of PTM structural-functional relationships.

6.5 Protein Ubiquitination

Ubiquitin (Ub) is a small protein, consisting of 76 amino acids, whose reversible covalent attachment to proteins governs such diverse biological processes as proteasomal degradation, DNA repair, activation of transcription factors, intracellular trafficking, and regulation of histones. Four different human genes are known to code for ubiquitin: UBB and UBC, which both code for polyubiquitin chains, and UBA52 and RPS27C, which code for single copies of ubiquitin fused to L40 and S27A, respectively.

In addition to Ub, several other small, ubiquitin-like proteins (Ubls) can be similarly attached to the lysine residue of a target protein. These Ubls include three isoforms of small ubiquitin-related modifier (SUMO1, SUMO2, SUMO3), neural precursor cell expressed, developmentally down-regulated 8 (NEDD8), ubiquitin-like protein ISG15 (ISG15), ubiquitin D (UBD), ubiquitin-like 5 (UBL5), ubiquitin-related modifier 1 homolog (URM1), autophagy associated protein 8 (ATG8), and ubiquitin-like protein ATG12 (ATG12). The focus of this discussion will be Ub, but occasional comparisons will be made to other Ubls.

Ub is attached to proteins through its C-terminal glycine residue via the epsilon amino group of a Lys residue in the target protein, forming an amide bond, through a series of enzymatic reactions (Fig. 6.4). In the first of these reactions, Ub is bound by an Ub-activating enzyme (E1), along with Mg^{2+} and ATP. E1 then catalyzes C-terminal acyl adenylation of the Ub chain, activating it for further reaction with a cysteine sulfhydryl group of E1. The activated Ub is then transferred to an Ub-conjugating enzyme, E2, through a transthioesterification reaction. Attachment to the target protein occurs through ubiquitin ligases, E3. The attachment of Ubls follows the same pathway, with characteristic E1/E2/E3 enzymes to catalyze the process. For both Ub and Ubls, the species attached can either be monomeric or polymeric, and the nature of the attachment helps signal the functional role of the modification.

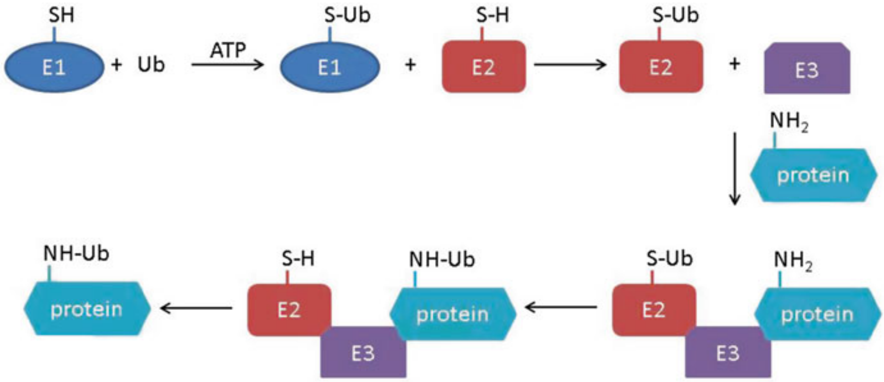


Fig. 6.4 The synthetic pathway involved in protein ubiquitination. Ub is attached to a Cys sulfhydryl residue of a Ub-activating enzyme (E1) in a process that is catalyzed by Mg^{2+} and ATP. The activated Ub is then transferred to the Cys sulfhydryl group of an Ub-conjugating enzyme (E2). Attachment to the target protein is mediated by a ubiquitin ligase (E3), which recruits both the target protein and an Ub-charged E2

Ubiquitin itself has seven Lys residues (at positions 6, 11, 27, 29, 33, 48 and 63) in addition to the N-terminus, and all provide potential sites for the Ub chain to be expanded. The signal that is sent by polyubiquitination is determined by which of these residues is modified. PolyUb chains can be linear, formed through end-to-end linkages via the N-terminus, or they can be branched through attachment at any of the other Lys residues. Attachment of at least four Ubs at Lys 48 is a known signal for proteasomal degradation (Voutsadakis 2007), and Ub chains at Lys 6 and Lys 11 have also been seen as proteasomal degradation signals (Voutsadakis 2012). Lys 63 ubitquitination can be a signal for autophagy, DNA repair or receptor kinase endocytosis (Jadhav and Wooten 2009).

To date, there are eight known E1 enzymes for activation of Ub and Ubls. (Schulman and Harper 2009). Interestingly, there are two known E1s for ubiquitin (UBA1 and UBA6), while a single E1 (SAE1-UBA2 heterodimer) is responsible for the activation of all SUMOs. Although each E1 recognizes Ub or a particular class of Ubl, all share a common mechanism of action. Key to the action is the adenylation domain, which can be homo- or heterodimeric in various members of the E1 family. In addition to catalyzing the acyl adenylation reaction, this domain's key function is to recognize the Ub or Ubl whose reaction it is catalyzing. Once the Ub, ATP and Mg^{2+} are positioned in the active site, the C-terminus is acyl adenylated, liberating inorganic phosphate. The Ub acyl adenylate is then attacked by a cysteine sulfhydryl group from the catalytic cysteine domain, resulting in formation of a new thioester linkage between Ub and the cysteine sulfhydryl group. Once this happens, a second Ub is then acyl adenylated and bound in the adenylation domain. The role of this second Ub is not known but is postulated to play a role in stabilizing the active conformation of the enzyme for transferring Ub between E1 and E2. Formation of the E1-Ub thioester triggers association of the appropriate E2 enzyme and subsequent Ub transfer. The ubiquitin fold domain (UFD) plays a key role in recognizing and binding to the correct E2, and a conformational change in the UFD is necessary to bring the E1-thioubiquitin close to the reactive E2 cysteine residue so that transthioesterification can occur.

Approximately forty ubiquitin E2s are encoded in the human genome, and all possess a highly conserved region of 150–200 amino acids that is known as the ubiquitin-conjugating catalytic (UBC) fold (van Wijk and Timmers 2010). The UBC is of central importance in the function of E2s as it is involved in binding E1s, E3s, and the activated Ub. Classification of E2s is determined by the presence or absence of additional sequence modification to the UBC. Class I E2s contain only the UBC fold, while Class II and III E2s contain additional sequence on the N- and C-termini, respectively. Class IV E2s are modified at both the N- and C-termini. The differences in sequence between the four classes of E2s help to define their differences in subcellular localization, interaction with E1s and E3s, and modulation of the activity of an interacting E3 (van Wijk and Timmers 2010). However, the main determinants for E2 properties are found within key regions of the UBC. An antiparallel beta sheet and one alpha helix (H2) form a central region, bounded by alpha helix 1 (H1) on one side and alpha helices 3 and 4 (H3 and H4) on the other.

Substrate specificity and recruitment are mediated by E3 ubiquitin ligases, and the fact that there are ~600 putative E3 ligases encoded into the human genome provides clues regarding the diversity of substrates for ubiquitination. As with E2s, E3s can be either monomeric, homodimeric or heterodimeric. E3s fall into two broad classes, based upon the domain responsible for binding to E2. Most human E3s contain a RING (Really Interesting New Gene) domain (Metzger et al. 2014). Two Zn^{2+} ions are bound in the RING domain, providing a scaffold for binding E2. A small subset of this first class of E3s, the RING-like U-box domain family, binds E2 in a similar manner but without the requirement for Zn^{2+} ions. The RING and RING-like E3s mediate the transfer of ubiquitin from E2 to the substrate of interest without themselves being modified by ubiquitin, instead facilitating the process by bringing the reactive species in close proximity to one another. In contrast, E3s in the second category are modified by ubiquitin in an intermediate catalytic step. This family of E3s contains a HECT (Homologous to E6-AP Carboxy Terminus) domain, and a conserved cysteine residue serves to transfer Ub to the substrate via formation of an intermediate E3-Ub thioester (Scheffner and Kumar 2014). In the case of the HECT family of EC ligases, Ub or Ubl specificity is found within the catalytic HECT domain. For RING family E3s, the Ub or Ubl whose attachment is being catalyzed depends upon the specific combination of E2 and E3. In other words, RING family E3s can generate different Ub linkages depending upon the E2 to which it is interacting. For HECT family E3s, protein substrate specificity is governed by a region that is located on the N-terminal side of the HECT domain.

Factors governing substrate recognition are not completely understood, but some trends are beginning to emerge (Jadhav and Wooten 2009). For instance, it is thought that cytosolic E3s recognize misfolded proteins through long stretches of hydrophobic residues. Other signals may include post-translational modifications or primary sequence. In terms of primary sequence, proteins with N-terminal Phe, Leu, Asp, Lys, or Arg residues were found to have a very short half-life (2–3 min.) when compared to amino acids with "stabilizing" the N-terminal residues Met, Ser, Ala, Thr, Val and Gly (Bachmair et al. 1986). This is known as the N-end rule. The destabilizing amino acids plus the Lys residue to be modified with Ub are referred to as an N-degron; they are recognized by E3s called N-recognins and targeted for destruction by the 26S proteasome (Jadhav and Wooten 2009). Other important primary sequence signals for ubiquitination include PEST sequences, so called because the sequences are rich in the amino acids Pro, Glu, Ser and Thr (Rogers et al. 1986); D-box, containing the consensus sequence R-A/T-A-L-G-X-I/V-G/T-N (Glotzer et al. 1991); and the KEN (Lys-Glu-Asn) box domain, with the consensus sequence K-E-N-X-X-X-N (Pfleger and Kirschner 2000). Post-translational modifications, including phosphorylation and glycosylation, have also been shown to activate a substrate toward ubiquitination.

As is the case with many post-translational modifications, ubiquitination is a reversible reaction. The proteolytic cleavage of ubiquitin side chains is catalyzed by a class of enzymes known as deubiquitinases (DUBs). At present, there are approximately 79 known DUBs (Komander et al. 2009) which follow several mechanistic pathways depending upon the polyubiquitin bond being proteolytically cleaved. As previously mentioned, ubiquitin can be transcribed as a polyUb chain or as fusion protein, so a DUB is crucial in producing monoUb. Also, once proteins have been targeted for degradation, the Ub and polyUb chains can be removed by DUBs for the purpose of recycling Ub and maintaining a steady state intracellular Ub concentration. DUBs can also remove Ub and polyUb in order to reverse Ub signaling and rescue proteins from degradation. There are five families of DUBs, and all possess a Ub-binding domain (UBD). Most DUBs catalyze cleavage of the bond between the epsilon-amino group of Lys and the C-terminus of Ub.

As is the case with many PTMs, the substoichiometric nature of ubiquitination can lead to challenges in identification of modified proteins. An additional challenge lies in the fact that ubiquitinated proteins are typically targeted for degradation. Therefore, enrichment strategies are a key component in the global mapping of ubiquitination. One of the earliest reported strategies for global study of ubiquitination involved a cell culture study of a cell line expressing a His-tagged form of Ub. A Ni-NTA column was used to isolate proteins modified by Ub. Proteolytic digestion with trypsin and LC-MS/MS analysis yielded 4210 peptides corresponding to 1,075 candidate ubiquitin-conjugating proteins (Peng et al. 2003). Ub-binding domains, covalently attached to beads, have also been shown to be effective in the enrichment of Ub proteins (Nakayasu et al. 2013), as have linkage-specific polyubiquitin antibodies (Matsumoto et al. 2010; Newton et al. 2008).

When ubiquitinated peptides are proteolytically digested with trypsin, a Gly-Gly motif remains on ubiquitinated lysine residues. Also, due to the modification of the Lys, trypsin will not cleave at this residue. The mass of the Gly-Gly remnant is 114.04 Da, and database search engines can use this value to identify modified peptides. Some caution must be used, though, because this particular change in mass is also characteristic for the addition of two carbamidomethyl groups, formed from over-alkylation with iodoacetamide. The Gly-Gly tag is also formed upon cleavage of two Ub-like proteins (NEDD8 and ISG15), so care must be taken when reporting Ub sites identified on the basis of a Gly-Gly residue.

Antibodies have been developed for the Lys-epsilon-Gly-Gly (K-ε-GG) residue, and the use of this antibody has been shown to be a viable strategy for enrichment and identification of ubiquitinated peptides. Most recently, the K-ε-GG antibody was used to study the effects of proteasomal and DUB inhibition in Jurkat cells, and 5533 K-ε-GG peptides were identified (Udeshi et al. 2012). Another recent report identified diagnostic b_2' and a_1' fragment ions for K-ε-GG peptides which were labeled with formaldehyde-D2 and $NaCNBH_3$ (Chicooree et al., 2013). These ions could prove to be valuable in the development of targeted MS-based strategies for the identification of ubiquitinated peptides.

6.6 Lipid Modifications of Proteins

The three major types of lipid modifications of proteins include covalent modifications by fatty acids (N-myristoylation, S- and N-palmytoylation), isoprenoids, and glycosylphosphatidyl inositol (GPI). These modifications may occur separately or in combinations, e.g. myristate + palmitate, palmitate + cholesterol, or farnesyl + palmitate. Some less frequent types of protein lipidation have been also described, such as cholesterol esterification of Hedgehog (Hh) proteins occurring in the ER during autoprocessing (Porter et al. 1996; Ryan and Chiang 2012; Grover et al. 2011; Palm et al. 2013), or direct attachment of a phospholipid phosphatidylethanolamine via an amide bond to the yeast protein Atg8 (Ichimura et al. 2000; Kirisako et al. 2000) and its mammalian homolog LC3 (Kabeya et al. 2004) during autophagy; however, the currently known examples are restricted to small groups of proteins.

The covalent linkage between a protein and either thioester-linked palmitate or a GPI anchor can be broken by the actions of thioesterases: cytosolic acyl protein thioesterase 1 (APT1) (Duncan and Gilman 1998; Zeidman et al. 2009) and lysosomal palmitoyl-protein thioesterase 1 (PPT1) (Camp and Hofmann 1993; Camp et al. 1994), and phospholipases (Low and Prasad 1988; Davitz et al. 1989; Metz et al. 1994), respectively. By contrast, neither myristate nor the isoprenoids farnesyl or geranylgeranyl are physically removed from a modified protein. Instead, some proteins sequester these lipophilic groups within a hydrophobic cleft, effectively shielding them from the aqueous milieu (Zozulya and Stryer 1992).

Until recently detection and analysis of lipid-modified proteins by traditional proteomic approaches has been relatively difficult and was mostly confined to studying individual purified proteins. Most methods relied on metabolic incorporation of radioactively labeled precursors into the proteins with subsequent detection, which required long exposure times, typically 1–3 months. This, as well as relatively high cost, use of hazardous reagents, and relatively low sensitivity made such methods unsuitable for high-throughput analysis of lipid-modified proteins on a proteome scale (Resh 2006; Martin et al. 2008). Application of mass-spectrometry for studying lipid modifications was also focused on individual lipidated proteins due to their relatively low representation and abundance in the total cellular proteome. The conventional separation techniques used for characterizing the global cellular proteome are not ideally suited for the recovery of highly hydrophobic lipidated peptides (Tom and Martin 2013). This problem may be partially overcome by optimizing the fractionation strategies (Ujihara et al. 2008; Serebryakova et al. 2011; Wotske et al. 2012). However, the major advance in the field came with the introduction of selective chemical labeling of the lipid modification sites.

The first global characterization of protein palmitoylation was performed in yeast (Roth et al. 2006) using acyl biotin exchange (ABE) enrichment technique, which involves alkylation of free thiol groups by N-ethylmaleimide or methyl methanethiosulfonate with subsequent cleavage of thioesters by hydroxylamine (HA) and labeling the exposed cysteine residues with a biotin analogue for further detection or affinity enrichment (Drisdel and Green 2004). This approach combined with multidimensional liquid chromatography (multidimensional protein identification technology (MudPIT) (Washburn et al. 2001) has allowed identification of 47 proteins, including 12 of the 14 previously known. Using deletion mutants for each of the seven yeast DHCC genes and their combinations, the authors found a significant overlap in substrate specificity between the different DHCC isoforms (Roth et al. 2006). This observation was later confirmed by Emmer and co-authors for *Trypanosoma brucei*. Using the same approach, they have identified a total of 124 palmitoylated proteins with an estimated false discovery rate of 1.0 % (Emmer et al. 2011). Kang and co-authors used a combination of ABE enrichment, MudPit and semi-quantitative analysis by Western blotting to study palmitoylated neuronal proteins; they identified the majority of the previously known proteins, as well as a significant number of novel putative protein substrates of palmitoylation (113 candidates in the high confidence group and 318 in the lower confidence group). Palmitoylation has been tested and confirmed for 21 of the newly identified proteins (Kang et al. 2008). A similar strategy was used to explore the protein palmitoylome and analyze DHHC substrate specificity in human endothelial cells (Marin et al. 2012) and macrophages (Merrick et al. 2011). The ABE approach has been also applied to characterize the distribution of palmitoylated proteins between the lipid rafts and non-raft membrane domains in a prostate cancer cell line. The authors applied a combined strategy for palmitoyl protein identification and site characterization (PalmPISC) which involved parallel SDS-PAGE separation and in-gel digestion of individual streptavidin-captured proteins and analysis of biotinylated peptides purified after in-solution digestion of the total protein extract. They have identified 67 known and 331 novel candidate S-acylated proteins, as well as the localization of 25 known and 143 novel candidate S-acylation sites (Yang et al. 2010). The PalmPISC strategy was also used to analyze protein palmitoylation in human platelets (Dowal et al. 2011).

All of these studies relied on spectral counting for relative quantification of proteins present in the experimental (HA-treated) and control (mock-treated) samples to identify proteins specifically enriched by biotin tag capture, with an arbitrary cutoff for the ratio of spectral counts under HA+versus HA- conditions. Samples were normalized by the spectral counts of co-purifying contaminants present in both types of samples; however, their content between samples may vary, potentially leading to inaccurate estimates. Zhang and co-authors used a quantitative ABE approach with differential isotope-coded affinity labeling by biotinylated, thiol-reactive isotope-coded affinity tag (ICAT) reagents to identify the protein substrates of the DHHC2 palmitoyltransferase in HeLa cells. The ratios of ABE-enriched proteins from cells with DHHC2 knockdown and control cells, labeled (H) and light (L) ICAT reagents, were calculated using combined peak areas for mutually shared isotopic masses (Zhang et al. 2008).

Quantitative comparison performed with 4-plex isobaric tags for relative and absolute quantitation (iTRAQ) (Hemsley et al. 2013) allowed quantitative comparison between the samples from wild-type *Arabidopsis* and a *tip1-2* mutant deficient for one of the palmitoyltransferases. This method facilitated reliable identification of a total of 561 putative S-acylated proteins and 103 candidate protein targets of the TIP1 palmitoyltransferase.

The acyl-biotin exchange and similar techniques have significantly advanced the proteomic characterization of palmitoylated proteins; they are relatively simple and inexpensive. However, this approach has certain inherent limitations and pitfalls. While it is well suited for studying S-acylated proteins due to the chemical lability of the thioester bond, it cannot be applied for labeling the sites of N-myristoylation, since the lipid moiety is attached to proteins through the stable amide bond. The specificity of palmitoylated protein identification is strongly dependent on the efficiency of the alkylation for complete blockage of all free thiol groups; high numbers of false-positive initial protein identifications which were discarded on the statistical basis could be explained by incomplete alkylation prior to the affinity enrichment step (Emmer et al. 2011; Merrick et al. 2011). Enzymes which use thioester-linked acyl intermediates in their reaction mechanism are another source of false-positive identifications (Yang et al. 2010). An alternative chemical approach to labeling and enrichment of lipid-modified proteins relies on metabolic labeling by bioorthogonal analogs of fatty acids or isoprenoids which contain a reactive group, usually a terminal alkyne or azide, which is used for fluorescent tagging or affinity labeling. This principle was first applied to the analysis of protein farnesylation in COS-1 cells by using an azido-farnesyl analog and subsequently conjugating it through the azide group to biotinylated phosphine capture reagent (bPPCR) using Staudinger ligation reaction. Affinity purification and proteomic analysis of the conjugated proteins led to the identification of 18 farnesylated proteins (Kho et al. 2004). The same chemical approach was used to identify palmitoylated proteins by metabolic incorporation of synthetic azido-tetradecanoic acid, an isosteric analog of palmitate (Kostiuk et al. 2008), as well as for labeling of geranylgeranylated proteins using an azido-geranylgeranyl analog with subsequent fluorescent tagging and detection (Chan et al. 2009).

Another type of bioorthogonal probe, an isosteric analog of palmitate with an ω-terminal acetylene group, 17-octadecynoic acid (17-ODYA), was used for profiling protein palmitoylation in Jurkat T cells, yielding identification of approximately 125 predicted palmitoylated proteins belonging to different functional groups. Like azido-tetradecanoic acid, it was efficiently incorporated by cellular palmitoylation machinery into the protein substrates, but the fluorescent and affinity tags for detection and enrichment were linked through the Cu(I)-catalyzed azide-alkyne Huisgen cycloaddition reaction (Martin and Cravatt 2009).

Bio-orthogonal approaches allow temporal control of probe incorporation for pulse–chase analysis. Since they do not require thiol reduction and alkylation and multiple precipitation steps, they are more suitable for analyzing smaller samples. Their overall specificity for the targeted lipid modifications is also higher, since it does not depend on thorough modification and shielding of all the free thiol groups in the analyzed proteins; however, false-positive identification of lipid-modified proteins is not totally excluded (Martin 2013). A combined strategy of identification of palmitoylated proteins in *Plasmodium falciparum,* in which the ABE approach and bio-orthogonal metabolic labeling were applied in parallel experiments proved useful for cross-validation of the resulting hits (Jones et al. 2012).

6.7 Inference of PTMs from MS Data

PTMs play an essential role in the protein's destiny and its function. However, it is challenging to identify them in the complex samples, even if the additional enrichment steps are applied. Widely-used search engines such as Mascot (Perkins et al. 1999), X!Tandem (Craig and Beavis 2003), OMSSA (Geer et al. 2004) or SEQUEST (Eng et al. 1994) can routinely identify only a restricted number of user predefined modifications. Standard database search algorithms employ protein amino acid sequences for *in silico* digestion, according to protease cleavage rules, into peptide fragments. Theoretical spectra are matched with empirical spectra and similarities are calculated. Both fixed and variable modifications are specified by the user prior to the search, forcing the search engines to align spectra with the mass shift of the modification (Potthast et al. 2007; Savitski et al. 2011; Bandeira et al. 2007). Typically, oxidation of methionine is specified as a variable modification. However, the inclusion of increased numbers of variable modifications causes search space to expand exponentially. This leads to longer processing times and a greater number of false positive assignments. Thus, platforms developed for the unrestricted identification of modified peptides execute a simple workflow (Ahrne et al. 2010) that can be described as: extraction of candidate peptides/proteins, matching the theoretical spectra and probability assignment (in general, mimicking the standard database search for the peptide identifications). Examples of such software are InsPecT (Tanner et al. 2005), Popitam (Hernandez et al. 2003), P-Mod (Hansen et al. 2005), VEMS 3.0 (Matthiesen et al. 2005), ModifiComb (Savitski et al. 2006), OpenSea (Searle et al. 2005).

The first step of the pipeline is focused on the reduction of the target database to improve matching scores. Most algorithms employ multiple round processing and sequence tag extraction, which can be used separately or in combination. During multiple rounds processing, the data is searched in two rounds to discard peptides which cannot be identified with sufficient confidence. At first, very strict rules are applied, allowing for one missed cleavage and one or two variable modifications. In round two, the list of significantly identified proteins from the previous step is screened again but with greater tolerance for PTMs. The Bonanza algorithm (Falkner et al. 2008) uses a similar approach, but is applied to search a spectral library. The algorithm assumes that unmodified and modified peptides have the same fragmentation patterns. Sequence tag extraction is based on a *de novo* sequencing algorithm (Dancik et al. 1999; Fernandez-de-Cossio et al. 2000; Johnson and Taylor 2002) to identify a three-four amino acid long peptide "tag" (Mann and Wilm 1994). InsPecT (Tanner et al. 2005) implements such algorithm followed by a trie-based scan of a database. Another way to improve the filtering process is to split the spectrum into intervals and extract the candidate peptides sequentially from each of them based on the reference database, so-called SIMS (Liu et al. 2008). The benefit of the database reduction step is that it decreases the search time and it can be performed externally. For instance, Swiss Protein Identification Toolbox, SwissPIT (Quandt et al. 2009), combines multiple identification tools to create a concise protein database which is then transferred to a PTM search engine such as Popitam or InsPecT.

In the matching step, a theoretical spectrum of *b*- and *y*-ions is generated for each candidate peptide. Usually, the counts for shared peaks between compared spectra are their similarity measure (Craig and Beavis 2003; Geer et al. 2004). Spectral libraries, on the other hand, contain experimental data which include information on the peak intensities. The use of spectral libraries therefore provides better scoring discriminants. QuickMod was designed as a tool for modification spectral library search (Ahrne et al. 2011). However, most algorithms function on a simple assumption of the similarity of the fragmentation patterns, which is not valid in some cases. The mass of the glutamic acid is the same as the mass of the methylated aspartic acid, which would generate the same theoretical spectrum. For this reason, some tools refer to the modification databases such as Unimod (Creasy and Cottrell 2004), DeltaMass (http://www.abrf.org/index.cfm/dm.home) or RESID (Garavelli 2004). Some of the modifications cause neutral losses during fragmentation or produce a diagnostic ion which could be considered during the PTM assignment and could help to distinguish between such modifications as lysine acetylation and lysine trimethylation. This algorithm is implemented in VEMS 3.0.

The last step of the standard PTM search algorithm is designed to refine the results. The approach proposed by Tsur et al. (2005) is based on the reasonable assumption that erroneous modification assignments would be distributed randomly throughout the dataset when scanning through all possible mass shifts for each amino acid. Therefore, only peptides with the modifications assigned to an amino acid reported multiple times are considered true positives. In the simplest case, a peptide contains only one possible site of modification but often there are several that could be modified. For instance, methionine residues are less common amino acids and are usually present only once in the peptide, making methionine oxidation site assignment unnecessary. In contrast, when anticipating for proline and tryptophan oxidation as well, the number of potential sites increases dramatically, making the differentiation between these residues more important. While the standard database search yields reasonable results in defining whether a peptide is modified or not, confident modification site localization may not be delivered. Currently, there are a number of commercial and in-house designed algorithms addressing this issue. Some of these tools are fully integrated into available MS/MS search engines, whereas others are independent isolated platforms performing only the modification site localization.

Generally, all the tools can be divided in two major groups: the ones that estimate the chance of a given peak to be matched randomly or calculate the score reflecting the difference between peptide identifications with various site localizations. The most known and probably the earliest tool in this field was A-Score (Beausoleil et al. 2006) for the SEQUEST search engine, originally designed to handle low resolution data. This algorithm calculates the probability of a site assignment based on the number of unique peaks for distinguishing between two possible sites *b*- and *y*-ions within a 100 Da window, which is then $\log_{10}$ transformed and multiplied by −10. A similar approach is implemented in InsPecT, SloMo (Albuquerque et al. 2008), and Phosphinator (Phanstiel et al. 2011). PTM Score, developed for the Andromeda search engine, uses the same algorithm but the scores are calculated for each potential site of the peptide and assuming that peptide is modified, the total probability of the modification and therefore all the scores is 100 % (Olsen et al. 2006). A bottleneck in these methods is the erroneous assumption that the chance to match a given mass difference is equally random. However, there are masses that match various amino acids with different modification, and there are masses that can't be matched to any combination. The PhosphoRS scoring algorithm addresses this issue by enabling the extraction of different numbers of peaks, from different spectral areas, within a defined (100 Da) *m/z* window, using the user-defined mass tolerance (Taus et al. 2011). Therefore, this tool is appropriate for the analysis of data obtained in a high mass accuracy instrument.

Other types of scoring algorithms are Mascot Delta Score (Savitski et al. 2011), the SLIP score within Protein Prospector (Baker et al. 2011), and variable modification localization scoring in Spectrum Mill (Agilent). These algorithms calculate the site localization reliability based on the difference between protein identification probabilities. In standard database searches, all potential modification sites are considered and the probability of the correct peptide assignment is estimated. The log10 transformed difference between those values defines the score for each site. However, Spectrum Mill uses the number of matched peaks, their type and the relative intensity of unmatched peaks to estimate the score. The major difference between the three methods is the number of extracted peaks. The SLIP score and Spectrum Mill use fixed number of peaks per spectrum (40 and 25 respectively) whereas the number of peaks extracted with Mascot Delta Score varies to keep peptide identification confidence at the optimal level. Additionally, both the SLIP scoring tool and Spectrum Mill are implemented within the corresponding search engines, enabling routine analysis of site modification localization.

However, the results still require verification because existing tools are not 100 % efficient. The amount of data produced with the current technology in shotgun proteomics is too large to be manually validated, pushing the development of tools and algorithms to estimate, analogous to the false discovery rate (FDR) for peptide identification measures, false localization rate (FLR). FDR measures are calculated based on the ratio between the number of false positives identified in a decoy database to the total number of identified peptides. It is assumed that their matching frequencies are the same. Unfortunately, it is nearly impossible to have similar estimates for the FLR, because the localization sites have to be known *a priori* to have an exact measure. Thus, there is insufficient information to calculate FLR for all the potential modification sites in a specific dataset. Additionally, identification of the peptide with an erroneously assigned site is not a random match; in contrast, it is a similar event to the correct assignment, which makes the use of decoy sequences an inefficient approach. The latest advance in this field is the Batch-Tag search engine that implements FLR estimation using amino acids that biologically could not be modified (Baker et al. 2010). During the decoy search, glutamate and proline residues were allowed to be phosphorylated. Proline phosphorylation does not occur in nature and phosphorylated glutamate exists only for a very short period of time as intermediate during the biosynthesis of glutamate and proline, making it unlikely to be detected in proteomic studies. In this way, all modifications assigned to proline or glutamate will be incorrect, enabling the estimation of the FLR. There is no direct correlation between a high peptide match score and FLR, because peptides with high scores could still be found to have high FLRs. Also, there is a higher risk of incorrect site assignment when there are two potential sites locating close to each other.

While in the case of phosphorylation the situation is more or less clear, in order to fully understand the signaling mechanisms and cellular responses comprehensive analysis of the other modification types is essential (Wang et al. 2007). Even though suitability claims for general PTM assignments for the previously described algorithms have been made, they are best suited for the phosphopeptide analysis.

In the case of ubiquitination, more advanced methods are required to discern between ubiquitin and ubiquitin-like peptide modifiers. Because the modification (Ub/Ubl) is a protein by nature, it is digested and fragmented during the MS/MS analysis, making spectral interpretation difficult. Spectra produced by Ub-modified peptides include *b*- and *y*-ions of the target peptide and *b*- and *y*-ions of the Ub/Ubl itself. SUMOylation pattern recognition tools may be used to identify peptide modifiers, as in SUMmOn, which considers only the most intense peaks within a 100 Da window (Pedrioli et al. 2006). The algorithm by Kang *et al.* is applicable for unrestricted PTM identification as well as Ub/Ubl modifier (Kang and Yi 2011). The algorithm consists of four stages of PTM identification and two stages of peak matching of Ub/Ubl *b*-ions with the measured peaks, and matching Ub/Ubl *y*-ions with mass shift classes. The differences between all measured peaks and theoretical fragment ions are calculated and divided into mass shift classes which are then filtered based on their intensity, mass deviation, and the number of mass differences in the class. Usually, Ub/Ubl identification relies on Gly-Gly or Leu-Arg-Gly-Gly mass shift on Lys (but not Ub *y*-ions such as Gly and Arg-Gly-Gly) and *b*-ions of free (not attached to target peptide) Ub/Ubls are mainly used for that. However, cysteines alkylated with iodoacetomide during sample preparation have the same mass as Gly-Gly, causing erroneous identification of Ub/Ubls (Jeram et al. 2009; Witze et al. 2007). In order to generate the sequence of attached Ub/Ubls the algorithm builds multiple mass shift paths based on matched mass shift classes and mass shifts of theoretical Ub/Ubl *y*-ions. All known and putative Ub/Ubl proteins used to evaluate the program were identified with 91 % accuracy when anticipating only for 1 PTM and 53 % for 2 PTMs.

Protein modification by glycosylation is important to consider. However, the complexity of the structure is significantly higher due to the branched nature of glycans, with different linkages and site isomers/isobars which differ only in their stereochemistry. The presence of such complex structures in the samples complicates the data search dramatically. There are few strategies to address this issue. Library-based sequencing tools (GlycosidIQ and SimGlycan) (Joshi et al. 2004; Apte and Meitei 2010), generate theoretical spectra for each glycan structure in the library and then match them to the measured spectrum, providing a score. Several approaches have emerged to process MS^n tandem mass spectrometry data. The saccharide topology analysis tool STAT (Gaucher et al. 2000) compares the list of all plausible oligosaccharide moieties for a predefined *m/z*, charge and product ion mass with the experimental spectrum, and provides the evaluation of the match. Similarly, Oscar (Lapadula et al. 2005), StrOligo (Ethier et al. 2003), and GlycoFragment (Lohmann and von der Lieth 2003), generate candidate structures from the predefined precursor ion and estimated composition but apply biosynthetic rule restrictions. Another way to perform the search is to match the spectra against a spectral library of oligosaccharides (Kameyama et al. 2005; Zhang et al. 2005). A *de novo*-based sequencing tool, GLYCH (Tang et al. 2005), allows the tree structure of a number of monosaccharide residues, maximizing the number of theoretical ions. Various structural solutions are then evaluated and ranked, taking in consideration one and two stages of fragmentation.

There is a demand for high quality empirical databases as well as technical infrastructure for glycomics. GlycomeDB (Ranzinger et al. 2011), GlycoSuiteDB (Cooper et al. 2001b; 2003), EUROCarbDB (von der Lieth et al. 2011), SWEET-DB (Loss et al. 2002), BOLD (Cooper et al. 1999), and KEGG (Hashimoto et al. 2006) are widely known and often used for glycan searches. GlycoSuiteDB is now included in UniCarbKB (http://www.unicarbkb.org/), GlycoWorkbench was designed by EUROCardDB initiative to evaluate manual spectrum annotation using the same approach as described above (Ceroni et al. 2008). For easier and faster structural assembly, it contains an intuitive visual editor, GlycanBuilder (Ceroni et al. 2007).

A completely different method to analyze glycans, a combinatorial approach, may be employed in glycan studies. GlycoMod (Cooper et al. 2001a) and Glycopeakfinder (Goldberg et al. 2005) allow *de novo* assignment of glycan composition from a single mass measurement. No prior knowledge of the biological background or fragmentation technique required. However, the number of possible compositions matching certain mass increases exponentially with the number of allowed monomers, leading to the development of tools that consider taxonomic and glycobiological background, such as Cartoonist (Goldberg et al. 2005) and Retrosynthetic Glycan Network Libraries (Kronewitter et al. 2009). Cartoonist, designed to analyze

MALDI-MS data, generates all plausible mammalian-synthesized N-linked glycan topologies using a manually compiled library of archetypes.

Despite the large variety of tools developed for glycoanalysis, there is little focus on raw spectral processing. The Glycolyzer (Kronewitter et al. 2012), designed on the basis of SysBioWare (Vakhrushev et al. 2009), is an integrated annotation program for glycan biomarker discovery. It contains all the basic components (background subtraction, peak detection, noise removal and data processing) as well as calibration, theoretical retrosynthetic library based glycan annotation and statistical hypothesis testing. The workflow uses FT-ICR MS data for the input.

The previously described bioinformatic tools are suitable for the oligosaccharide structural assignments. However, Byonic (Bern et al. 2012) and GlycoPeptideSearch (Pompach et al. 2012) combine the known glycan analysis methods and the proteomics search or user-specified potential glycosylated peptides. GlycoSearch (Kletter et al. 2013) is a related tool that is used for glycan binding motif analysis of lectins.

Finally, we will describe the development of bioinformatics tools for the analysis of lipid modifications of proteins. Sites of modification may be predicted *in silico* using numerous amino acid sequence-based tools for various lipids (partly available via http://mendel.imp.ac.at/). For proteomic purposes, lipoproteins are isolated from the sample and enriched narrowing the scope of the work for a standard peptide identification. However, for more information on structure, function, biosynthesis and association with certain protein (pathway) one could refer to LIPID Maps (http://www.lipidmaps.org/).

In spite of the rapid development of bioinformatics for automated identification and site localization of modifications, verification is still mostly manual. For that purpose, knowledge-based libraries of all modifications such as DeltaMass and UniMod are very useful, as well as PhosphoSitePlus (http://www.phosphosite.org/), Phospho.ELM (http://phospho.elm.eu.org/), PHOSIDA (Gnad et al. 2011) and METLIN (a metabolite database; http://metlin.scripps.edu) (Smith et al. 2005). However, in order to be able to identify novel modifications, *de novo* sequencing algorithms implemented in PepNovo (Frank and Pevzner 2005) or PEAKS (Ma et al. 2003) would be quite useful in this respect.

6.8 Summary

The large number of types PTMs that have been identified create an enormous challenge to proteomic studies modifications. The challenges are related to the chemical nature of PTMs, their microheterogeneity, and site localization. In this chapter, we highlighted the biological importance of the most common types of PTMS, namely, protein acetylation, phosphorylation, glycosylation, ubiquitination, and lipidation. We also provided an overview of separation methods, mass spectrometric analysis, and recent developments in bioinformatic strategies to analyze PTMs on a proteomic scale.

References

Ahrne E, Muller M, Lisacek F. Unrestricted identification of modified proteins using MS/MS. Proteomics. 2010;10:671–86.

Ahrne E, Nikitin F, Lisacek F, Muller M. QuickMod: a tool for open modification spectrum library searches. J Proteome Res. 2011;10:2913–21.

Albuquerque CP, Smolka MB, Payne SH, Bafna V, Eng J, Zhou H. A multidimensional chromatography technology for in-depth phosphoproteome analysis. Mol Cell Proteomics. 2008;7:1389–96.

Alley WR, Mechref Y, Novotny MV. Characterization of glycopeptides by combining collision-induced dissociation and electron-transfer dissociation mass spectrometry data. Rapid Commun Mass Spectrom. 2009;23:161–70.

Alpert AJ. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. J Chromatogr. 1990;9:177–96.

Alpert AJ. Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. Anal Chem. 2007;80:62–76.

Alpert AJ. Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. Anal Chem. 2008;80:62–76.

An HJ, Peavy TR, Hedrick JL, Lebrilla CB. Determination of N-glycosylation sites and site heterogeneity in glycoproteins. Anal Chem. 2003;75:5628–37.

Apte A, Meitei NS. Bioinformatics in glycomics: glycan characterization with mass spectrometric data using SimGlycan. Methods Mol Biol. 2010;600:269–81.

Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochimica et Biophys Acta (BBA) Gen Subj. 1999;1473:4–8.

Atwood JA, Minning T, Ludolf F, Nuccio A, Weatherly DB, Alvarez-Manilla G, Tarleton R, Orlando R. Glycoproteomics of *Trypanosoma cruzi* trypomastigotes using subcellular fractionation, lectin affinity, and stable isotope labeling. J Proteome Res. 2006;5:3376–84.

Bachmair A, Finley D, Varshavsky A. In vivo half-life of a protein is a function of its amino-terminal residue. Science. 1986;234:179–86.

Baker PR, Medzihradszky KF, Chalkley RJ. Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state- and sequence-dependent scoring. Mol Cell Proteomics. 2010;9:1795–803.

Baker PR, Trinidad JC, Chalkley RJ. Modification site localization scoring integrated into a search engine. Mol Cell Proteomics. 2011;10.

Bandeira N, Tsur D, Frank A, Pevzner P. Protein identification by spectral networks analysis. Proc Natl Acad Sci U S A. 2007;104:6140–5.

Bannister AJ, Miska EA, Görlich D, Kouzarides T. Acetylation of importin-α nuclear import factors by CBP/p300. Curr Biol. 2000;10:467–70.

Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006;24:1285–92.

Bern M, Kil YJ, Ecker C. Byonic: advanced peptide and protein identification software. Curr Protoc Bioinformatics. 2012. Chapter 13, Unit 13 20.

Boersema P, Mohammed S, Heck AR. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. Anal Bioanal Chem. 2008;391:151–9.

Bunkenborg J, Pilch BJ, Podtelejnikov AV, Wisniewski JR. Screening for N-glycosylated proteins by liquid chromatography mass spectrometry. Proteomics. 2004;4:454–65.

Camp LA, Hofmann SL. Purification and properties of a palmitoyl-protein thioesterase that cleaves palmitate from H-Ras. J Biol Chem. 1993;268:22566–74.

Camp LA, Verkruyse LA, Afendis SJ, Slaughter CA, Hofmann SL. Molecular cloning and expression of palmitoyl-protein thioesterase. J Biol Chem. 1994;269:23212–19.

Černý M, Skalák J, Cerna H, Brzobohatý B. Advances in purification and separation of posttranslationally modified proteins. J Proteomics. 2013;92:2–27.

Ceroni A, Dell A, Haslam SM. The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. Sour Code Biol Med. 2007;2:3.

Ceroni A, Maass K, Geyer H, Geyer R, Dell A, Haslam SM. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. J Proteome Res. 2008;7:1650–9.

Chan LN, Hart C, Guo L, Nyberg T, Davies BS, Fong LG, Young SG, Agnew BJ, Tamanoi F. A novel approach to tag and identify geranylgeranylated proteins. Electrophoresis. 2009;30:3598–606.

Chicooree N, Connolly Y, Tan CT, Malliri A, Li Y, Smith DL, Griffiths JR. Enhanced detection of ubiquitin isopeptides using reductive methylation. J Am Soc Mass Spectrom. 2013;24(3): 421–30.

Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

Cooper CA, Wilkins MR, Williams KL, Packer NH. BOLD–a biological O-linked glycan database. Electrophoresis. 1999;20:3589–98.

Cooper CA, Gasteiger E, Packer NH. GlycoMod–a software tool for determining glycosylation compositions from mass spectrometric data. Proteomics. 2001a;1:340–9.

Cooper CA, Harrison MJ, Wilkins MR, Packer NH. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. Nucleic Acids Res. 2001b;29:332–5.

Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. Nucleic Acids Res. 2003;31:511–13.

Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom. 2003;17:2310–16.

Creasy DM, Cottrell JS. Unimod: protein modifications for mass spectrometry. Proteomics. 2004;4:1534–6.

Crimmins DL, Gorka J, Thoma RS, Schwartz BD. Peptide characterization with a sulfoethyl aspartamide column. J Chromatogr. 1988;443:63–71.

Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. J Comput Biol. 1999;6:327–42.

Davies M, Smith KD, Harbin AM, Hounsell EF. High-performance liquid chromatography of oligosaccharide alditols and glycopeptides on a graphitized carbon column. J Chromatogr A. 1992;609:125–31.

Davitz MA, Hom J, Schenkman S. Purification of a glycosyl-phosphatidylinositol-specific phospholipase D from human plasma. J Biol Chem. 1989;264:13760–4.

Dell A, Morris HR. Glycoprotein structure determination by mass spectrometry. Science. 2001;291:2351–6.

Dephoure N, Gould KL, Gygi SP, Kellogg DR. Mapping and analysis of phosphorylation sites: a quick guide for cell biologists. Mol Biol Cell. 2013;24:535–42.

Di Palma S, Hennrich ML, Heck AJR, Mohammed S. Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. J Proteomics. 2012;75:3791–813.

Ding W, Hill JJ, Kelly J. Selective enrichment of glycopeptides from glycoprotein digests using ion-pairing normal-phase liquid chromatography. Anal Chem. 2007;79:8891–9.

Dormeyer W, Mohammed S, Breukelen BV, Krijgsveld J, Heck AJR. Targeted analysis of protein termini. J Proteome Res. 2007;6:4634–45.

Dowal L, Yang W, Freeman MR, Steen H, Flaumenhaft R. Proteomic analysis of palmitoylated platelet proteins. Blood. 2011;118:e62–73.

Drisdel RC, Green WN. Labeling and quantifying sites of protein palmitoylation. Biotechniques. 2004;36:276–85.

Dube DH, Bertozzi CR. Glycans in cancer and inflammation [mdash] potential for therapeutics and diagnostics. Nat Rev Drug Discov. 2005;4:477–88.

Duncan JA, Gilman AG. A cytoplasmic acyl-protein thioesterase that removes palmitate from G protein alpha subunits and p21(RAS). J Biol Chem. 1998;273:15830–7.

Eberharter A, Becker PB. Histone acetylation: a switch between repressive and permissive chromatin. EMBO Rep. 2002;3:224–9.

Edwards AVG, Edwards GJ, Schwämmle V, Saxtorph H, Larsen MR. Spatial and temporal effects in protein post-translational modification distributions in the developing mouse brain. J Proteome Res. 2014;13(1):260–7.

Emmer BT, Nakayasu ES, Souther C, Choi H, Sobreira TJ, Epting CL, Nesvizhskii AI, Almeida IC, Engman DM. Global analysis of protein palmitoylation in African trypanosomes. Eukaryot Cell. 2011;10:455–63.

Eng JK, Mccormack AL, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J Am Soc Mass Spectrom. 1994;5:976–89.

Ethier M, Saba JA, Spearman M, Krokhin O, Butler M, Ens W, Standing KG, Perrault H. Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. Rapid Commun Mass Spectrom. 2003;17:2713–20.

Eyrich B, Sickmann A, Zahedi RP. Catch me if you can: mass spectrometry-based phosphoproteomics and quantification strategies. Proteomics. 2011;11:554–70.

Falkner JA, Falkner JW, Yocum AK, Andrews PC. A spectral clustering approach to MS/MS identification of post-translational modifications. J Proteome Res. 2008;7:4614–22.

Fan JQ, Kondo A, Kato I, Lee YC. High-performance liquid chromatography of glycopeptides and oligosaccharides on graphitized carbon columns. Anal Biochem. 1994;219:224–9.

Ferluga S, Hantgan R, Goldgur Y, Himanen JP, Nikolov DB, Debinski W. Biological and structural characterization of glycosylation on ephrin-A1, a preferred ligand for EphA2 receptor tyrosine kinase. J Biol Chem. 2013;288:18448–57.

Fernandez-De-Cossio J, Gonzalez J, Satomi Y, Shima T, Okumura N, Besada V, Betancourt L, Padron G, Shimonishi Y, Takao T. Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. Electrophoresis. 2000;21:1694–9.

Ficarro SB, Adelmant GO, Tomar MN, Zhang Y, Cheng VJ, Marto JA. Magnetic bead processor for rapid evaluation and optimization of parameters for phosphopeptide enrichment. Anal Chem. 2009;81:4566–75.

Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem. 2005;77:964–73.

Garavelli JS. The RESID database of protein modifications as a resource and annotation tool. Proteomics. 2004;4:1527–33.

Gates M, Tomer K, Deterding L. Comparison of metal and metal oxide media for phosphopeptide enrichment prior to mass spectrometric analyses. J Am Soc Mass Spectrom. 2010;21:1649–59.

Gaucher SP, Cancilla MT, Phillips NJ, Gibson BW, Leary JA. Mass spectral characterization of lipooligosaccharides from Haemophilus influenzae 2019. Biochemistry. 2000;39:12406–14.

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res. 2004;3:958–64.

Gilar M, Olivova P, Daly AE, Gebler JC. Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. J Sep Sci. 2005;28:1694–703.

Gilar M, Yu Y-Q, Ahn J, Fournier J, Gebler JC. Mixed-mode chromatography for fractionation of peptides, phosphopeptides, and sialylated glycopeptides. J Chromatogr A. 2008;1191:162–70.

Glotzer M, Murray AW, Kirschner MW. Cyclin is degraded by the ubiquitin pathway. Nature. 1991;349:132–8.

Glozak MA, Sengupta N, Zhang X, Seto E. Acetylation and deacetylation of non-histone proteins. Gene. 2005;363:15–23.

Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res. 2011;39:D253–60.

Goldberg D, Sutton-Smith M, Paulson J, Dell A. Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. Proteomics. 2005;5:865–75.

Grover VK, Valadez JG, Bowman AB, Cooper MK. Lipid modifications of Sonic hedgehog ligand dictate cellular reception and signal response. PLoS One. 2011;6:e21353.

Gupta S, Seth A, Davis RJ. Transactivation of gene expression by Myc is inhibited by mutation at the phosphorylation sites Thr-58 and Ser-62. Proc Natl Acad Sci. 1993;90:3216–20.

Hägglund P, Bunkenborg J, Elortza F, Jensen ON, Roepstorff P. A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. J Proteome Res. 2004;3:556–66.

Hägglund P, Matthiesen R, Elortza F, Højrup P, Roepstorff P, Jensen ON, Bunkenborg J. An enzymatic deglycosylation scheme enabling identification of core fucosylated N-glycans and O-glycosylation site mapping of human plasma proteins. J Proteome Res. 2007;6:3021–31.

Hakansson K, Cooper HJ, Emmett MR, Costello CE, Marshall AG, Nilsson CL. Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptic to yield complementary sequence information. Anal Chem. 2001;73:4530–6.

Hansen BT, Davey SW, Ham AJL, Liebler DC. P-mod: an algorithm and software to map modifications to peptide sequences using tandem MS data. J Proteome Res. 2005;4:358–68.

Hao P, Guo T, Sze SK. Simultaneous analysis of proteome, phospho- and glycoproteome of rat kidney tissue with electrostatic repulsion hydrophilic interaction chromatography. PLoS One. 2011;6:e16884.

Hart GW, Slawson C, Ramirez-Correa G, Lagerlof O. Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. Annu Rev Biochem. 2011;80:825–58.

Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M. KEGG as a glycome informatics resource. Glycobiology. 2006;5:63R–70.

Heidinger-Pauli JM, Ünal E, Koshland D. Distinct targets of the Eco1 acetyltransferase modulate cohesion in S phase and in response to DNA damage. Mol Cell. 2009;34:311–21.

Hemsley PA, Weimar T, Lilley KS, Dupree P, Grierson CS. A proteomic approach identifies many novel palmitoylated proteins in Arabidopsis. New Phytol. 2013;197:805–14.

Hernandez P, Gras R, Frey J, Appel RD. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics. 2003;3:870–8.

Himanen JP, Goldgur Y, Miao H, Myshkin E, Guo H, Buck M, Nguyen M, Rajashankar KR, Wang B, Nikolov DB. Ligand recognition by A-class Eph receptors: crystal structures of the EphA2 ligand-binding domain and the EphA2/ephrin-A1 complex. EMBO Rep. 2009;10:722–8.

Himanen JP, Yermekbayeva L, Janes PW, Walker JR, Xu K, Atapattu L, Rajashankar KR, Mensinga A, Lackmann M, Nikolov DB, Dhe-Paganon S. Architecture of Eph receptor clusters. Proc Natl Acad Sci. 2010;107:10860–5.

Horn DM, Ge Y, Mclafferty FW. Activated ion electron capture dissociation for mass spectral sequencing of larger (42 KDa) proteins. Anal Chem. 2000;72:4778–84.

Hua S, Hu CY, Kim BJ, Totten SM, Oh MJ, Yun N, Nwosu CC, Yoo JS, Lebrilla CB, An HJ. Glyco-Analytical Multispecific Proteolysis (Glyco-AMP): a simple method for detailed and quantitative glycoproteomic characterization. J Proteome Res. 2013;12:4414–23.

Hwang C-S, Shemorry A, Varshavsky A. N-terminal acetylation of cellular proteins creates specific degradation signals. Science. 2010;327:973–7.

Ichimura Y, Kirisako T, Takao T, Satomi Y, Shimonishi Y, Ishihara N, Mizushima N, Tanida I, Kominami E, Ohsumi M, Noda T, Ohsumi Y. A ubiquitin-like system mediates protein lipidation. Nature. 2000;408:488–92.

Imperiali B, Rickert KW. Conformational implications of asparagine-linked glycosylation. Proc Natl Acad Sci. 1995;92:97–101.

Jadhav T, Wooten MW. Defining an embedded code for protein ubiquitination. J Proteomics Bioinform. 2009;2:316.

Jeram SM, Srikumar T, Pedrioli PG, Raught B. Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites. Proteomics. 2009;9:922–34.

Johnson RS, Taylor JA. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. Mol Biotechnol. 2002;22:301–15.

Jones ML, Collins MO, Goulding D, Choudhary JS, Rayner JC. Analysis of protein palmitoylation reveals a pervasive role in Plasmodium development and pathogenesis. Cell Host Microbe. 2012;12:246–58.

Joshi HJ, Harrison MJ, Schulz BL, Cooper CA, Packer NH, Karlsson NG. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. Proteomics. 2004;4:1650–64.

Kabeya Y, Mizushima N, Yamamoto A, Oshitani-Okamoto S, Ohsumi Y, Yoshimori T. LC3, GABARAP and GATE16 localize to autophagosomal membrane depending on form-II formation. J Cell Sci. 2004;117:2805–12.

Kameyama A, Kikuchi N, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Takahashi K, Narimatsu H. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. Anal Chem. 2005;77:4719–25.

Kang C, Yi GS. Identification of ubiquitin/ubiquitin-like protein modification from tandem mass spectra with various PTMs. BMC Bioinform. 2011;12 Suppl 14:S8.

Kang R, Wan J, Arstikaitis P, Takahashi H, Huang K, Bailey AO, Thompson JX, Roth AF, Drisdel RC, Mastro R, Green WN, Yates JR, Davis 3rd NG, El-Husseini A. Neural palmitoyl-proteomics reveals dynamic synaptic palmitoylation. Nature. 2008;456:904–9.

Kelleher NL, Zubarev RA, Bush K, Furie B, Furie BC, Mclafferty FW, Walsh CT. Localization of labile posttranslational modifications by electron capture dissociation: the case of gamma-carboxyglutamic acid. Anal Chem. 1999;71:4250–3.

Kho Y, Kim SC, Jiang C, Barma D, Kwon SW, Cheng J, Jaunbergs J, Weinbaum C, Tamanoi F, Falck J, Zhao Y. A tagging-via-substrate technology for detection and proteomics of farnesylated proteins. Proc Natl Acad Sci U S A. 2004;101:12479–84.

Kimura A, Matsubara K, Horikoshi M. A decade of histone acetylation: marking eukaryotic chromosomes with specific codes. J Biochem. 2005;138:647–62.

Kirisako T, Ichimura Y, Okada H, Kabeya Y, Mizushima N, Yoshimori T, Ohsumi M, TAKAO T, Noda T, Ohsumi Y. The reversible modification regulates the membrane-binding state of Apg8/Aut7 essential for autophagy and the cytoplasm to vacuole targeting pathway. J Cell Biol. 2000;151:263–76.

Kletter D, Cao Z, Bern M, Haab B. Determining lectin specificity from glycan array data using motif segregation and GlycoSearch software. Curr Protoc Chem Biol. 2013;5:157–69.

Koizumi K, Okada Y, Fukuda M. High-performance liquid chromatography of mono- and oligosaccharides on a graphitized carbon column. Carbohydr Res. 1991;215:67–80.

Komander D, Clague MJ, Urbe S. Breaking the chains: structure and function of the deubiquitinases. Nat Rev Mol Cell Biol. 2009;10:550–63.

Kostiuk MA, Corvi MM, Keller BO, Plummer G, Prescher JA, Hangauer MJ, Bertozzi CR, Rajaiah G, Falck JR, Berthiaume LG. Identification of palmitoylated mitochondrial proteins using a bio-orthogonal azido-palmitate analogue. FASEB J. 2008;22:721–32.

Kouzarides T. Acetylation: a regulatory modification to rival phosphorylation? EMBO J. 2000;19:1176–9.

Kronewitter SR, An HJ, de Leoz ML, Lebrilla CB, Miyamoto S, Leiserowitz GS. The development of retrosynthetic glycan libraries to profile and classify the human serum N-linked glycome. Proteomics. 2009;9:2986–94.

Kronewitter SR, DE Leoz ML, Strum JS, An HJ, Dimapasoc LM, Guerrero A, Miyamoto S, Lebrilla CB, Leiserowitz GS. The glycolyzer: automated glycan annotation software for high performance mass spectrometry and its application to ovarian cancer glycan biomarker discovery. Proteomics. 2012;12:2523–38.

Lapadula AJ, Hatcher PJ, Hanneman AJ, Ashline DJ, Zhang H, Reinhold VN. Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. Anal Chem. 2005;77:6271–9.

Larsen MR, Højrup P, Roepstorff P. Characterization of gel-separated glycoproteins using Two-step proteolytic digestion combined with sequential microcolumns and mass spectrometry. Mol Cell Proteomics. 2005;4:107–19.

Larsen MR, Jensen SS, Jakobsen LA, Heegaard NHH. Exploring the sialiome using titanium dioxide chromatography and mass spectrometry. Mol Cell Proteomics. 2007;6:1778–87.

Lewandrowski U, Zahedi RP, Moebius J, Walter U, Sickmann A. Enhanced N-glycosylation site analysis of sialoglycopeptides by strong cation exchange prefractionation applied to platelet plasma membranes. Mol Cell Proteomics. 2007;6(11):1933–41.

Lin R, Zhou X, Huang W, Zhao D, Lu L, Xiong Y, Guan KL, Lei QY. Acetylation control of cancer metabolism. Curr Pharm Des. 2014;20(15):2627–33.

Little DP, Speir JP, Senko MW, O'connor PB, Mclafferty FW. Infrared multiphoton dissociation of large multiply-charged ions for biomolecule sequencing. Anal Chem. 1994;66:2809–15.

Liu J, Erassov A, Halina P, Canete M, Nguyen DV, Chung C, Cagney G, Ignatchenko A, Fong V, Emili A. Sequential interval motif search: unrestricted database surveys of global MS/MS data sets for detection of putative post-translational modifications. Anal Chem. 2008;80:7846–54.

Lohmann KK, von der Lieth CW. GLYCO-FRAGMENT: a web tool to support the interpretation of mass spectra of complex carbohydrates. Proteomics. 2003;3:2028–35.

Loss A, Bunsmann P, Bohne A, Schwarzer E, Lang E, von der Lieth CW. SWEET-DB: an attempt to create annotated data collections for carbohydrates. Nucleic Acids Res. 2002;30:405–8.

Low MG, Prasad AR. A phospholipase D specific for the phosphatidylinositol anchor of cell-surface proteins is abundant in plasma. Proc Natl Acad Sci U S A. 1988;85:980–4.

Luo J, Su F, Chen D, Shiloh A, Gu W. Deacetylation of p53 modulates its effect on cell growth and apoptosis. Nature. 2000;408:377–81.

Luo J, Li M, Tang Y, Laszkowska M, Roeder RG, Gu W. Acetylation of p53 augments its site-specific DNA binding both in vitro and in vivo. Proc Natl Acad Sci U S A. 2004;101:2259–64.

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 2003;17:2337–42.

Malm D, Nilssen O. Alpha-mannosidosis. Orphanet J Rare Dis. 2008;3:21.

Mann M, Wilm M. Error tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem. 1994;66:4390–9.

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298:1912–34.

Manning JM, Popowicz AM, Padovan JC, Chait BT, Manning LR. Intrinsic regulation of hemoglobin expression by variable subunit interface strengths. FEBS J. 2012;279:361–9.

Marin EP, Derakhshan B, Lam TT, Davalos A, Sessa WC. Endothelial cell palmitoylproteomic identifies novel lipid-modified targets and potential substrates for protein acyl transferases. Circ Res. 2012;110:1336–44.

Marshall AG, Hendrickson CL. High-resolution mass spectrometers. In: Yeung SY, Zare RN, editors. Annual review of analytical chemistry. Palo Alto: Annual Reviews; 2008.

Martin BR. Chemical approaches for profiling dynamic palmitoylation. Biochem Soc Trans. 2013;41:43–9.

Martin BR, Cravatt BF. Large-scale profiling of protein palmitoylation in mammalian cells. Nat Methods. 2009;6:135–8.

Martin DD, Vilas GL, Prescher JA, Rajaiah G, Falck JR, Bertozzi CR, Berthiaume LG. Rapid detection, discovery, and identification of post-translationally myristoylated proteins during apoptosis using a bio-orthogonal azidomyristate analog. FASEB J. 2008;22:797–806.

Martin-Rendon E, Blake DJ. Protein glycosylation in disease: new insights into the congenital muscular dystrophies. Trends Pharmacol Sci. 2003;24:178–83.

Matsumoto ML, Wickliffe KE, Dong KC, Yu C, Bosanac I, Bustos D, Phu L, Kirkpatrick DS, Hymowitz SG, Rape M, Kelley RF, Dixit VM. K11-linked polyubiquitination in cell cycle control revealed by a K11 linkage-specific antibody. Mol Cell. 2010;39:477–84.

Matthiesen R, Trelle MB, Hojrup P, Bunkenborg J, Jensen ON. VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. J Proteome Res. 2005;4:2338–47.

Merrick BA, Dhungana S, Williams JG, Aloor JJ, Peddada S, Tomer KB, Fessler MB. Proteomic profiling of S-acylated macrophage proteins identifies a role for palmitoylation in mitochondrial targeting of phospholipid scramblase 3. Mol Cell Proteomics. 2011;10: M110.006007.

Metz CN, Brunner G, Choi-Muira NH, Nguyen H, Gabrilove J, Caras IW, Altszuler N, Rifkin DB, Wilson EL, Davitz MA. Release of GPI-anchored membrane proteins by a cell-associated GPI-specific phospholipase D. EMBO J. 1994;13:1741–51.

Metzger MB, Pruneda JN, Klevit RE, Weissman AM. RING-type E3 ligases: Master manipulators of E2 ubiquitin-conjugating enzymes and ubiquitination. Biochim Biophys Acta. 2014;1843(1):47–60.

Murphy M, Ahn J, Walker KK, Hoffman WH, Evans RM, Levine AJ, George DL. Transcriptional repression by wild-type p53 utilizes histone deacetylases, mediated by interaction with mSin3a. Genes Dev. 1999;13:2490–501.

Mysling S, Palmisano G, Højrup P, Thaysen-Andersen M. Utilizing ion-pairing hydrophilic interaction chromatography solid phase extraction for efficient glycopeptide enrichment in glycoproteomics. Anal Chem. 2010;82:5598–609.

Nakayasu ES, Ansong C, Brown JN, Yang F, Lopez-Ferrer D, Qian WJ, Smith RD, Adkins JN. Evaluation of selected binding domains for the analysis of ubiquitinated proteomes. J Am Soc Mass Spectrom. 2013;24:1214–23.

Newton K, Matsumoto ML, Wertz IE, Kirkpatrick DS, Lill JR, Tan J, Dugger D, Gordon N, Sidhu SS, Fellouse FA, Komuves L, French DM, Ferrando RE, Lam C, Compaan D, Yu C, Bosanac I, Hymowitz SG, Kelley RF, Dixit VM. Ubiquitin chain editing revealed by polyubiquitin linkage-specific antibodies. Cell. 2008;134:668–78.

Nilsson CL. Advances in quantitative phosphoproteomics. Anal Chem. 2011a;84:735–46.

Nilsson CL. Lectin techniques for glycoproteomics. Curr Proteomics. 2011b;8:248–56.

Nilsson IM, von Heijne G. Determination of the distance between the oligosaccharyltransferase active site and the endoplasmic reticulum membrane. J Biol Chem. 1993;268:5798–801.

Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006;127:635–48.

Oppermann FS, Gnad F, Olsen JV, Hornberger R, Greff Z, Keri G, Mann M, Daub H. Large-scale proteomics analysis of the human kinome. Mol Cell Proteomics. 2009;8:1751–64.

Packer NH, Lawson MA, Jardine DR, Redmond JW. A general approach to desalting oligosaccharides released from glycoproteins. Glycoconj J. 1998;15:737–47.

Palm W, Swierczynska MM, Kumari V, Ehrhart-Bornstein M, Bornstein SR, Eaton S. Secretion and signaling activities of lipoprotein-associated hedgehog and non-sterol-modified hedgehog in flies and mammals. PLoS Biol. 2013;11:e1001505.

Palmisano G, Lendal S, Larsen M. Titanium dioxide enrichment of sialic acid-containing glycopeptides. In: Gevaert K, Vandekerckhove J, editors. Gel-free proteomics. Totowa: Humana Press; 2011.

Palmisano G, Melo-Braga MN, Engholm-Keller K, Parker BL, Larsen MR. Chemical deamidation: a common pitfall in large-scale N-linked glycoproteomic mass spectrometry-based analyses. J Proteome Res. 2012;11:1949–57.

Pedrioli PG, Raught B, Zhang XD, Rogers R, Aitchison J, Matunis M, Aebersold R. Automated identification of SUMOylation sites using mass spectrometry and SUMmOn pattern recognition software. Nat Methods. 2006;3:533–9.

Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, Marsischky G, Roelofs J, Finley D, Gygi SP. A proteomics approach to understanding protein ubiquitination. Nat Biotechnol. 2003;21:921–6.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551–67.

Pfleger CM, Kirschner MW. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. Genes Dev. 2000;14:655–65.

Phanstiel DH, Brumbaugh J, Wenger CD, Tian SL, Probasco MD, Bailey DJ, Swaney DL, Tervo MA, Bolin JM, Ruotti V, Stewart R, Thomson JA, Coon JJ. Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat Methods. 2011;8:821–7.

Pompach P, Chandler KB, Lan R, Edwards N, Goldman R. Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. J Proteome Res. 2012;11:1728–40.

Porter JA, Young KE, Beachy PA. Cholesterol modification of hedgehog signaling proteins in animal development. Science. 1996;274:255–9.

Potthast F, Gerrits B, Hakkinen J, Rutishauser D, Ahrens CH, Roschitzki B, Baerenfaller K, Munton RP, Walther P, Gehrig P, Seif P, Seebergerg PH, Schlapbach R. The mass distance fingerprint: a statistical framework for de novo detection of predominant modifications using high-accuracy mass spectrometry. J Chromatogr B-Anal Technol Biomed Life Sci. 2007;854:173–82.

Quandt A, Masselot A, Hernandez P, Hernandez C, Maffioletti S, Appel RD, Lisacek F. SwissPIT: an workflow-based platform for analyzing tandem-MS spectra using the grid. Proteomics. 2009;9:2648–55.

Ranzinger R, Herget S, von der Lieth CW, Frank M. GlycomeDB–a unified database for carbohydrate structures. Nucleic Acids Res. 2011;39:D373–6.

Reich NC. STAT3 Revs up the powerhouse. Sci Signal. 2009;2:pe61.

Reinhold V, Zhang H, Hanneman A, Ashline D. Toward a platform for comprehensive glycan sequencing. Mol Cell Proteomics. 2013;12:866–73.

Resh MD. Use of analogs and inhibitors to study the functional significance of protein palmitoylation. Methods. 2006;40:191–7.

Rivera CM, Ren B. Mapping human epigenomes. Cell. 2013;155:39–55.

Robinson NE, Robinson A. Deamidation of human proteins. Proc Natl Acad Sci. 2001;98:12409–13.

Rogers S, Wells R, Rechsteiner M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science. 1986;234:364–8.

Roth AF, Wan J, Bailey AO, Sun B, Kuchar JA, Green WN, Phinney BS, Yates 3rd JR, Davis NG. Global analysis of protein palmitoylation in yeast. Cell. 2006;125:1003–13.

Rutishauser U. Polysialic acid in the plasticity of the developing and adult vertebrate nervous system. Nat Rev Neurosci. 2008;9:26–35.

Ryan KE, Chiang C. Hedgehog secretion and signal transduction in vertebrates. J Biol Chem. 2012;287:17905–13.

Satomi Y, Shimonishi Y, Takao T. N-glycosylation at Asn491 in the Asn-Xaa-Cys motif of human transferrin. FEBS Lett. 2004;576:51–6.

Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics. 2006;5:935–48.

Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, Kuster B. Confident phosphorylation site localization using the mascot delta score. Mol Cell Proteomics. 2011;10:M110.003830.

Schachter H, Freeze HH. Glycosylation diseases: Quo vadis? Biochimica et Biophys Acta (BBA) Mol Basis Dis. 2009;1792:925–30.

Scheffner M, Kumar S. Mammalian HECT ubiquitin-protein ligases: biological and pathophysiological aspects. Biochim Biophys Acta. 2014;1843(1):61–74.

Schulman BA, Harper JW. Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways. Nat Rev Mol Cell Biol. 2009;10:319–31.

Scott NE, Parker BL, Connolly AM, Paulech J, Edwards AVG, Crossett B, Falconer L, Kolarich D, Djordjevic SP, Højrup P, Packer NH, Larsen MR, Cordwell SJ. Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of *Campylobacter jejuni*. Mol Cell Proteomics. 2011;10.

Searle BC, Dasari S, Wilmarth PA, Turner M, Reddy AP, David LL, Nagalla SR. Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. J Proteome Res. 2005;4:546–54.

Serebryakova MV, Demina IA, Galyamina MA, Kondratov IG, Ladygina VG, Govorun VM. The acylation state of surface lipoproteins of mollicute Acholeplasma laidlawii. J Biol Chem. 2011;286:22769–76.

Singh C, Zampronio CG, Creese AJ, Cooper HJ. Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) mass spectrometry for the analysis of N-linked glycoproteins. J Proteome Res. 2012;11:4517–25.

Siuti N, Kelleher NL. Decoding protein modifications using top-down mass spectrometry. Nat Methods. 2007;4:817–21.

Smith CA, O'maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. METLIN: a metabolite mass spectral database. Ther Drug Monit. 2005;27:747–51.

Spange S, Wagner T, Heinzel T, Krämer OH. Acetylation of non-histone proteins modulates cellular signalling at multiple levels. Int J Biochem Cell Biol. 2009;41:185–98.

Starheim KK, Gromyko D, Evjenth R, Ryningen A, Varhaug JE, Lillehaug JR, Arnesen T. Knockdown of human Nα-terminal acetyltransferase complex C leads to p53-dependent apoptosis and aberrant human Arl8b localization. Mol Cell Biol. 2009;29:3569–81.

Sykes SM, Mellert HS, Holbert MA, Li K, Marmorstein R, Lane WS, Mcmahon SB. Acetylation of the p53 DNA-binding domain regulates apoptosis induction. Mol Cell. 2006;24:841–51.

Tang H, Mechref Y, Novotny MV. Automated interpretation of MS/MS spectra of oligosaccharides. Bioinformatics. 2005;21 Suppl 1:i431–9.

Tanner S, Shu HJ, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: identification of posttransiationally modified peptides from tandem mass spectra. Anal Chem. 2005;77:4626–39.

Taus T, Kocher T, Pichler P, Paschke C, Schmidt A, Henrich C, Mechtler K. Universal and confident phosphorylation site localization using phosphoRS. J Proteome Res. 2011;10:5354–62.

Tipton JD, Tran JC, Catherman AD, Ahlf DR, Durbin KR, Kelleher NL. Analysis of intact protein isoforms by mass spectrometry. J Biol Chem. 2011;286:25451–8.

Tom CT, Martin BR. Fat chance! Getting a grip on a slippery modification. ACS Chem Biol. 2013;8:46–57.

Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. Nat Biotechnol. 2005;23:1562–7.

Udeshi ND, Mani DR, Eisenhaure T, Mertins P, Jaffe JD, Clauser KR, Hacohen N, Carr SA. Methods for quantification of in vivo changes in protein ubiquitination following proteasome and deubiquitinase inhibition. Mol Cell Proteomics. 2012;11:148–59.

Ujihara T, Sakurai I, Mizusawa N, Wada H. A method for analyzing lipid-modified proteins with mass spectrometry. Anal Biochem. 2008;374:429–31.

Vakhrushev SY, Dadimov D, Peter-Katalinic J. Software platform for high-throughput glycomics. Anal Chem. 2009;81:3252–60.

van Damme P, Arnesen T, Gevaert K. Protein alpha-N-acetylation studied by N-terminomics. FEBS J. 2011;278:3822–34.

van Wijk SJ, Timmers HT. The family of ubiquitin-conjugating enzymes (E2s): deciding between life and death of proteins. FASEB J. 2010;24:981–93.

Varki A. Biological roles of oligosaccharides: all of the theories are correct. Glycobiology. 1993;3:97–130.

Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW, Etzler ME. Essentials of glycobiology. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2009.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, Mckusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, DI Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji R-R, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J-L, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, salzberg S, Shao W, Shue B, Sun J, Wang ZY, Wang A, Wang X, Wang J, Wei M-H, Wides R, Xiao C, Yan C, et al. The sequence of the human genome. Science. 2001;291:1304–51.

von der Lieth CW, Freire AA, Blank D, Campbell MP, Ceroni A, Damerell DR, Dell A, Dwek RA, Ernst B, Fogh R, Frank M, Geyer H, Geyer R, Harrison MJ, Henrick K, Herget S, Hull WE, Ionides J, Joshi HJ, Kamerling JP, Leeflang BR, Lutteke T, Lundborg M, Maass K, Merry A, Ranzinger R, rosen J, Royle L, Rudd PM, Schloissnig S, Stenutz R, Vranken WF, Widmalm G, Haslam SM. EUROCarbDB: an open-access platform for glycoinformatics. Glycobiology. 2011;21:493–502.

Voutsadakis IA. Pathogenesis of colorectal carcinoma and therapeutic implications: the roles of the ubiquitin-proteasome system and Cox-2. J Cell Mol Med. 2007;11:252–85.

Voutsadakis IA. Ubiquitination and the ubiquitin-proteasome system as regulators of transcription and transcription factors in epithelial mesenchymal transition of cancer. Tumour Biol. 2012;33:897–910.

Wang W, Yang X, Kawai T, De Silanes IL, Mazan-Mamczarz K, Chen P, Chook YM, Quensel C, Köhler M, Gorospe M. AMP-activated protein kinase-regulated phosphorylation and acetylation of importin α1: involvement in the nuclear import of RNA-binding protein HuR. J Biol Chem. 2004;279:48376–88.

Wang Z, Pandey A, Hart GW. Dynamic interplay between O-linked N-acetylglucosaminylation and glycogen synthase kinase-3-dependent phosphorylation. Mol Cell Proteomics. 2007;6:1365–79.

Washburn MP, Wolters D, Yates 3rd JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol. 2001;19:242–7.

Wimley WC, Creamer TP, White SH. Solvation energies of amino acid side chains and backbone in a family of host–guest pentapeptides†. Biochemistry. 1996;35:5109–24.

Witze ES, Old WM, Resing KA, Ahn NG. Mapping protein post-translational modifications with mass spectrometry. Nat Methods. 2007;4:798–806.

Woodworth A, Fiete D, Baenziger JU. Spatial and temporal regulation of tenascin-R glycosylation in the cerebellum. J Biol Chem. 2002;277:50941–7.

Wotske M, WU Y, Wolters DA. Liquid chromatographic analysis and mass spectrometric identification of farnesylated peptides. Anal Chem. 2012;84:6848–55.

Wu S-L, Huhmer AFR, Hao Z, Karger BL. On-line LC–MS approach combining collision-induced dissociation (CID), electron-transfer dissociation (ETD), and CID of an isolated charge-reduced species for the trace-level characterization of proteins with post-translational modifications. J Proteome Res. 2007;6:4230–44.

Wuhrer M, Koeleman CAM, Hokke CH, Deelder AM. Protein glycosylation analyzed by normal-phase nano-liquid chromatography–mass spectrometry of glycopeptides. Anal Chem. 2004;77:886–94.

Yang XJ, Seto E. HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. Oncogene. 2007;26:5310–18.

Yang WH, Kim JE, Nam HW, Ju JW, Kim HS, Kim YS, Cho JW. Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. Nat Cell Biol. 2006;8:1074–83.

Yang W, Di Vizio D, Kirchner M, Steen H, Freeman MR. Proteome scale characterization of human S-acylated proteins in lipid raft-enriched and non-raft membranes. Mol Cell Proteomics. 2010;9:54–70.

Yuan H, Marmorstein R. Histone acetyltransferases: rising ancient counterparts to protein kinases. Biopolymers. 2013;99:98–111.

Zeidan Q, Hart GW. The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. J Cell Sci. 2010;123:13–22.

Zeidman R, Jackson CS, Magee AI. Protein acyl thioesterases (Review). Mol Membr Biol. 2009;26:32–41.

Zhang H, Aebersold R. Isolation of glycoproteins and identification of their N-linked glycosylation sites. In: New and emerging proteomic techniques. Totowa: Humana Press; 2006.

Zhang H, Li X, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nat Biotechnol. 2003;21:660–6.

Zhang H, Singh S, Reinhold VN. Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. Anal Chem. 2005;77:6263–70.

Zhang J, Planey SL, Ceballos C, Stevens Jr SM, Keay SK, Zacharias DA. Identification of CKAP4/p63 as a major substrate of the palmitoyl acyltransferase DHHC2, a putative tumor suppressor, using a novel proteomics method. Mol Cell Proteomics. 2008;7:1378–88.

Zhang H, Guo T, Li X, Datta A, Park JE, Yang J, Lim SK, Tam JP, Sze SK. Simultaneous characterization of glyco- and phosphoproteomes of mouse brain membrane proteome with electrostatic repulsion hydrophilic interaction chromatography. Mol Cell Proteomics. 2010;9:635–47.

Zozulya S, Stryer L. Calcium-myristoyl protein switch. Proc Natl Acad Sci U S A. 1992;89:11569–73.

Zubarev RA, Kelleher NL, Mclafferty FW. Electron capture dissociation of multiply charged protein cations. A nonergodic process. J Am Chem Soc. 1998;120:3265–6.

Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, Mclafferty FW. Electron capture dissociation for structural characterization of multiply charged protein cations. Anal Chem. 2000;72:563–73.

Chapter 7
Biomarker Discovery Utilizing Biobanking Archives and the Diagnostic Market

Melinda Rezeli, Karin Sjödin, David Erlinge, and György Marko-Varga

Abstract The challenges in everyday healthcare are very often related to the ability to provide the appropriate therapy to the patient. In this respect personalized treatment has became the focus of the modern medicine. The various forms of diseases require new diagnostic, therapeutic and prognostic markers that can serve as targets for personalized drug development. Biobank establishment has a great importance globally, as biobank has been identified as a key area in the drug development and in the discovery of new protein biomarkers. The fast progression of biobanks around the world is becoming an important resource for society where the patient benefit is in the focus. As part of a Swedish national cardiological research initiative, the development of a quantitative MRM assay is reported for the quantification of putative cardiovascular disease markers. The assay was utilized for the analysis of patient samples taken from the LUNDHEARTGENE biobank.

Keywords Cardiovascular disease • MRM • Protein sequencing • Proteomics • Biobank • Mass spectrometry

M. Rezeli • K. Sjödin
Clinical Protein Science & Imaging, Biomedical Center, Biomedical Engineering,
Lund University, BMC D13, 221 84 Lund, Sweden

Center of Excellence in Biological and Medical Mass Spectrometry,
Biomedical Center D13, Lund University, 221 84 Lund, Sweden

D. Erlinge
Department of Cardiology, Lund University, Skåne University Hospital,
221 85 Lund, Sweden

G. Marko-Varga (✉)
Clinical Protein Science and Imaging, Biomedical Center, Biomedical Engineering,
Lund University, BMC D13, 221 84 Lund, Sweden

Center of Excellence in Biological and Medical Mass Spectrometry,
Biomedical Center D13, Lund University, 221 84 Lund, Sweden

First Department of Surgery, Tokyo Medical University,
6-7-1 Nishishinjiku Shinjuku-ku, Tokyo 160-0023, Japan
e-mail: gyorgy.marko-varga@bme.lth.se

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_7

7.1 Introduction

The challenges in everyday healthcare are very often related to the ability to provide the right medication, at the right time to the right patient. In this respect patient stratification is becoming the real focus of modern drug strategies, with the aim to optimize so that when a drug is given to a patient it will result in a response to the given treatment. Targeted treatments by Personalized Medicine (PM), a successful 2nd generation drug that targets the key regulating protein in disease (Hamburg and Collins 2010; Marko-Varga et al. 2007) by blocking or partially inhibiting the activity of these protein targets, will have a direct impact on the disease onset. The future of biomedical sciences will be driven by the ability to adopt novel technologies, generating large datasets, to understand disease mechanisms and to develop new treatments. This is especially relevant to diseases such as cardiac infarct and vascular diseases (Anderson 2005; Vegvari and Marko-Varga 2010).

Today, imaging techniques used to diagnose organ disease status, include ultrasonography (US), computed tomography (CT) or magnetic resonance imaging (MRI). The development of better strategies for diagnosis and treatment are lagging behind other types of malignancies and to date no tissue or blood biomarker, gene signature, or molecular targets exist in cardiovascular diseases. Proteomics is a modern protein expression technology, capable of mapping and quantifying a whole spectrum of proteins (Zolg and Langen 2004; Rezeli et al. 2013). These study activities provide a bridge to relate patterns of disease-perturbed proteins with specific diseases by the use of quantitative proteomics and especially multiple reaction monitoring (MRM), that delivers a pattern of quantitative read-out (Hüttenhain et al. 2009; Végvári et al. 2013; Krastins et al. 2013). The development of new diagnostic biomarkers has a great potential, where both industry and academy are investing and searching for approaches to improve the discovery successes, where new technology plays a central role.

Due to the improved quality of life, the elderly group of society is rapidly increasing. Today, the elderly population in society is higher than 22 % in countries like Japan, Sweden, France, and is ever increasing expected to reach up to more than 30 % of national populations, in Japan even as high as 40 %. This will result in an addition of an increasing cost to society, to the expenses of the health care and governmental costs. This patient group will use their own funds to buy medicine or treatments in addition to the government. Today co-morbidity, which reflects a life of a patient, combines several diseases and gives more complexity to health care treatment and increases the cost of health care. Hospitals and health care institutions are searching for new technology platforms that can facilitate the establishment of the proper diagnosis and monitor the effectiveness of the given treatment. It is roughly estimated that 10 % of the health care sector is spent on diagnostics. The market is characterized by a high growth, especially in developing nations. The Indian market for example, fuelled by the arrival of a broad range of modern lifestyle illnesses including diabetes and cardiovascular diseases, its home market expand at a rate of 20 % per year. While an aging population requires more diagnostics, the health care system is keen on reining in costs. At the same time the industry is searching for innovations that offer faster and easier analyses (http://www.frost.com/prod/servlet/market-insight-top.pag?Src=RSS&docid=118961271).

Biobanking initiatives is a central part of healthcare activity that provides a major asset with patient materials that can be used for future developments, bringing patient treatments closer to curing, (Simenon-Dubach and Perren 2011; Eiseman et al. 2003). The biomarker concept is not a new invention per se. Its utilization has been exploited in both medicine, as well as within the drug development process. However, a distinction should be made to the surrogate marker, that is a definite indicator of a medical, or pathophysiological event. The National Institutes of Health (NIH) Biomarker working group defined a biomarker as "*a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention*" (Biomarkers Definition Working Group 2001).

7.2 Biomarkers and Diagnostics

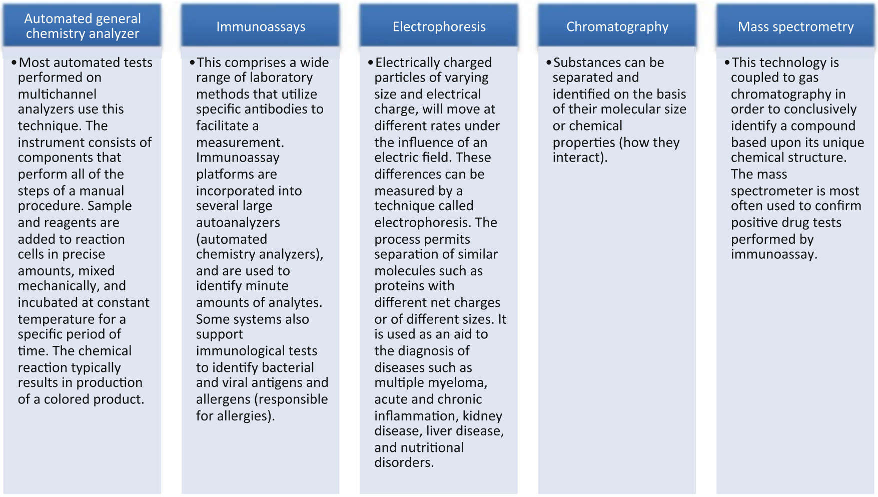
Today patients awareness about disease indication and disease management is rapidly increasing and the patient can also support his/her treatment financially in order to get better and faster treatment results. With this development at hand, new diagnostic products that are introduced into the healthcare sector on the market, will have high priority in society. Currently the healthcare budgets are not actually covering the total health care cost where the raising demand, develops into that more and more initiatives is developed into alternative solutions to cover costs.

The emerging market is already shifting from North America to Asian countries like China, Korea, India and Japan. A progressive growth of India's US35 billion healthcare industries can be seen in the rise of the number of diagnostic laboratories due to the increasing health awareness. Currently there are over 12,000 hospitals and 15,000 diagnostic laboratories contributing extensively to the overall healthcare delivery market in India. 70 % of medical decisions are based on the diagnostic tests, making this clearly a driving force. Medical devices supply market in India grew from US1.19 billion in 2008 to US1.70 billion in 2010. By 2012, it is estimated to reach US2.78 billion. India spends 5.1 % of GDP on healthcare (http://www.medindia.net/news/Growth-of-Clinical-Diagnostic-Laboratories-in-India-Fueled-by-Health-Awareness-69489-1.htm).

7.3 Clinical Diagnostics and Treatment for Research

The clinical diagnostics and research industry comprises of businesses and laboratories that offer analytic or diagnostic services including body fluid analysis and diagnostic imaging that aid the medical professions. Medical diagnosis is medical determination of disease or syndrome performed by a physician. The focus is on the disease process and the physical, genetic, or environmental cause of that process. The medical laboratory, also called the clinical laboratory or the pathology

Table 7.1 Outline of the various biomedical analysis technology areas



Automated general chemistry analyzer	Immunoassays	Electrophoresis	Chromatography	Mass spectrometry
• Most automated tests performed on multichannel analyzers use this technique. The instrument consists of components that perform all of the steps of a manual procedure. Sample and reagents are added to reaction cells in precise amounts, mixed mechanically, and incubated at constant temperature for a specific period of time. The chemical reaction typically results in production of a colored product.	• This comprises a wide range of laboratory methods that utilize specific antibodies to facilitate a measurement. Immunoassay platforms are incorporated into several large autoanalyzers (automated chemistry analyzers), and are used to identify minute amounts of analytes. Some systems also support immunological tests to identify bacterial and viral antigens and allergens (responsible for allergies).	• Electrically charged particles of varying size and electrical charge, will move at different rates under the influence of an electric field. These differences can be measured by a technique called electrophoresis. The process permits separation of similar molecules such as proteins with different net charges or of different sizes. It is used as an aid to the diagnosis of diseases such as multiple myeloma, acute and chronic inflammation, kidney disease, liver disease, and nutritional disorders.	• Substances can be separated and identified on the basis of their molecular size or chemical properties (how they interact).	• This technology is coupled to gas chromatography in order to conclusively identify a compound based upon its unique chemical structure. The mass spectrometer is most often used to confirm positive drug tests performed by immunoassay.

laboratory, provides diagnostic testing services for physicians to help identify the cause of disease and changes produced in the body by disease conditions. Medical laboratories are classified based on the type of services they render such as bacteriological, blood and other body fluid, pathology, mycology and, parasitology laboratories, which are all clinical units and are our future potential customers. Commercial medical laboratories operate as independent businesses and serve as testing facilities for physicians and for companies engaged in medical or pharmaceutical research. Medical laboratories depend upon computer-controlled automated equipment for as many tests as possible to keep up with the volume and variety of tests ordered. The goal of such automation is to reduce the amounts of sample required; reduce the amount of chemicals (reagents) needed per test; reduce the time of analysis; eliminate contamination and error that results from excessive sample handling; and reduce the number of technologists needed to perform the testing. Cost savings achieved through automation are important to both the testing facility and the patient. Table 7.1 comprises an overview of the most common Biomedical analysis technology platform areas.

7.4 Biobanking

A biobank is defined as a storage facility for long-term storage of human samples that are identifiable to a specific person and linked to personal data. As a new trend in healthcare activities that relate to epigenetics, and epidemiology, population-based research biobanks will collect, environmental and lifestyle samples and generated data that enables large scale meta data analysis. When Time magazine named Biobanking as one of the "10 Ideas Changing the World Right Now" (Park 2009).

One of the target areas where biobanking plays an important role is within the area of drug discovery, development and diagnostic developments. In this respect, modern drug discovery and drug development within the pharmaceutical industry is heavily dependent on biobank resources comprising a wealth of clinical patient materials (Hewitt 2011).

Extensive resources on a global level have been invested in population-based studies where the aim is to gather large cohorts of participants that gives a good representation of our society. Examples of these clinical activities are large-scale studies, such as the UK Biobank (http://www.ukbiobank.ac.uk/) and the LifeGene study (https://www.lifegene.se/In-english/). In these studies there is an absolute control of patient sampling with standardized procedures. Storage handlings are in most cases performed at −80 °C that provides the basis for high quality biobank samples (Malm et al. 2013). The novel technologies are currently changing the way that we perform standard procedures in hospitals. The most important variables such as temperature and sample preparations that are associated with sample instability have been taken into consideration with the development of these new technologies, in order to keep high quality of patient samples for long periods.

The organization, and coordination of samples that are associated with clinical data combines the value that can be obtained from these sample archives. By the implementation of e-health logistics, efficient data storage and use, allows a data history to be established where the patient treatments are linked to the decision made by the physician, providing healthcare improvements. A novel technology that relates to biomarkers and clinical status of the patients, is a great resource that utilizes blood samples as a major biofluid resource. Blood sample storages nowadays range from large national efforts into smaller development labs where the biospecimen collections are used as biobank assets, searching for healthcare solutions (LaBaer 2012; Marko-Varga et al. 2011). Such an example is the "The Swedish Web-system for Enhancement and Development of Evidence-based care in Heart disease Evaluated According to Recommended Therapies" (SWEDEHEART), (Jernberg et al. 2010).

A local initiative in south of Sweden is the Lund HeartGene biobank initiative.

One objective in these studies is the early identification of cardiac infarct. In addition, adverse cardiac remodeling that is following the myocardial infarction remains a significant cause of congestive heart failure, which utilizes major resources at the hospital. Our challenge is to improve our ability to make early diagnosis, and predict and treat early. There are several studies and research groups that have identified single and multiple candidate biomarkers and strategies that identifies diagnostic prognosticators of cardiac events (Kuhn et al. 2009; Addona et al. 2011; Domanski et al. 2012).

One candidate that has proven to have a close cardiac disease link is MMP-9 (matrix metalloproteinase 9). This target protein actually has been identified both as a drug target as well as a potential biomarker for cardiac remodeling. This was demonstrated with both animal models as well as within clinical studies. It is found that MMP-9 expression significantly increases and is linked with inflammation, diabetic microvascular complications, extracellular matrix degradation and synthesis, and cardiac dysfunction (Yabluchanskiy et al. 2013).

7.5 Biomarker Discovery

The analysis of proteins in a biological context is not novel and histopathologists, under the guise of immunohistochemistry, have been dabbling in protein studies for many years. However, this activity, like many 'classical' protein studies has profound limitations of analytical sensitivity, the inability to readily study functional correlates and, most importantly, the inability to put a protein of interest into the context of other proteins being produced by the cell. The enormous advances made in analytical chemistry have made possible the relatively new discipline of proteomics. Cardiovascular disease (CVD) is the leading cause of death in the high-income parts of the world (Perk et al. 2012) even though preventive measures and acute treatments have improved substantially in later years. In contrary to a basic understanding of the cardiovascular system, it is clear that in cardiovascular diseases, the heart is not able, itself receive enough oxygen and nutrients from the blood it pumps and it must be supplied with blood. Cardiac infarct disorders, and malfunctions of the coronary circulation can have devastating effects to the heart. As the injury to the heart can reduce coronary circulation, will continue to progress and give rise to further damages.

By definition, a proximal biomarker shows a close relationship with its target disease, whereas a distal biomarker exhibits non-targeted disease modifying outcomes. There is an unmet need of new biomarkers in the field, for improved early diagnosis and risk assessment of patients with cardiovascular disease. The availability of novel biomarkers for diagnostic and prognostic applications in healthcare is expanding as a result of improvements in proteomic methods.

The quantification within a MRM study may be broadly defined as a collection of scientific and technical approaches designed to characterise the protein content of cells, tissues and whole organisms. One objective is to find the mechanisms of protein expression and post-synthetic modification by comparing samples originating from patients in health and disease. As a subject, it has also come to mean the multi-parameter analysis of the protein products of a cell or tissue and, as such, has benefited hugely from developments in advanced analysis, e.g. in mass spectrometry and in associated informatics capabilities. Figure. 7.1 illustrates a common MS-based proteomics workflow, where the proteins are enzymatically digested and alternatively further processed through a number of sample preparation steps, and then high resolution liquid chromatographic separation is applied for the separation of the resulted peptide mixture interfaced with tandem mass spectrometry. Protein expression may be further categorised into the following activities: expression proteomics (the detection and analysis of the proteome using a range of analytical approaches); functional proteomics (the analysis of the function and regulation of proteins); structural proteomics (the determination and analysis of the three-dimensional structure of proteins) and chemi-proteomics (the study of the interaction of proteins with pharmacologically active, small molecular weight compounds).

As a part of a cardiology initiative study within the Swedish national cardiological research, our research team developed a quantitative MRM assay for putative cardiovascular disease markers. We utilized patient samples taken from the

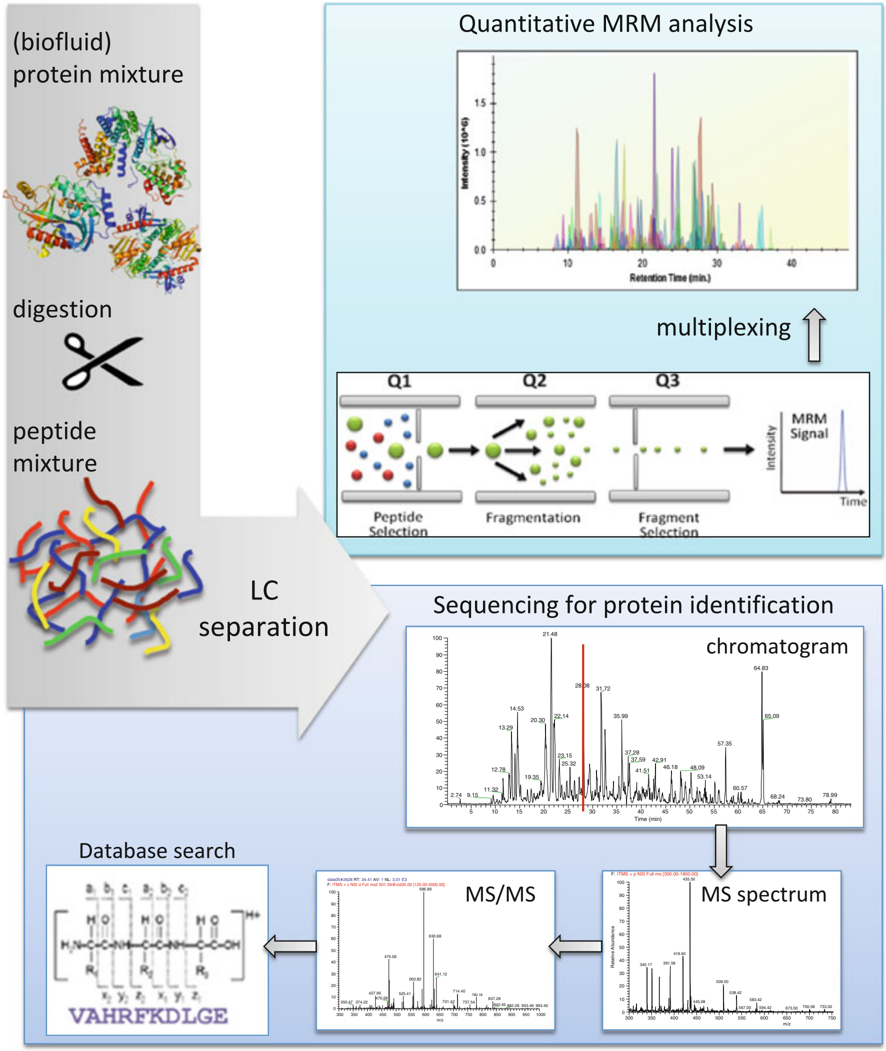


Fig. 7.1 MS-based proteomics workflow, where the proteins are digested and the resulting peptide mixture is analyzed by high-resolution liquid chromatography coupled with tandem mass spectrometry

LUNDHEARTGENE biobank. We processed these plasma samples by using an optimized digestion protocol followed by nanoLC–MRM/MS analysis.

This MRM assay was developed in order to generate a high throughput screening platform for simultaneous quantification of a large set of putative cardiovascular disease markers as target proteins, which was next applied to biobanking material.

The objective of this study is to make comparative analysis of patients with ST-segment elevation myocardial infarction in relation to patients suffering from chest pain that arises due to other causes.

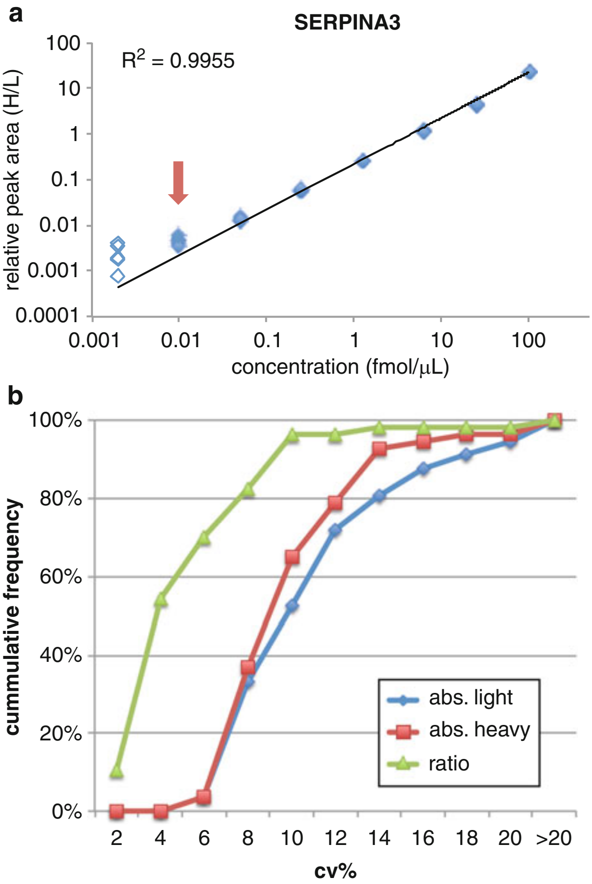


Fig. 7.2 (**a**) Standrad curve of SERPINA3 generated in pooled plasma digest. Linearity of the MRM assay determined by using heavy labeled IS peptides spiked into pooled plasma digest at various concentrations. The *arrow* indicates the LOQ (cv<20 %) **(b)** Intra-assay variability of MRM quantitation. CV frequencies of the MRM assay based on triplicate measurements of 28 proteins in spiked plasma digests

In this cardio biomarker study, where we quantify multi-protein expression, we outlined and developed the experimental conditions and assay parameters, that provides highly reproducible protein quantification (Fig. 7.2); in addition the sequence of each and every marker within the assay was confirmed by synthetic peptide standards and utilizing database search, as shown in Fig. 7.1.

The basic principles of the mass spectrometric analysis utilizing the triple quadrupole principle is depicted in Fig. 7.1. As shown in Fig. 7.2a, we obtain typically excellent linear regressions for the target peptides within 4-5 orders of magnitude linear range with LOQ ranged in the femtomolar level. Figure 7.2b provides data on intra-assay variability of MRM quantification of 28 protein targets. The graph illustrates the frequency of the coefficient of variation of the MRM assay based on triplicate analysis, and clearly shows the impact of the application of heavy internal standards on the assay reproducibility.

Acknowledgements We would like to acknowledge The Kamprad Foundation, Thermo, the Swedish Foundation for Strategic Research, "TOTAL-AMI", the Swedish Research Council, Vinnova, Knut and Alice Wallenberg Foundation, Crafoord Foundation, Carl Trygger Foundation and European Research Council Advanced Grant ERC-2011-294576.

References

Addona TA, et al. A pipeline that integrate the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. Nat Biotechnol. 2011;29:635–43.

Anderson NL. The role of multiple proteomic platforms in a pipeline for new diagnostics. Mol Cell Proteomics. 2005;4:1441–4.

Biomarkers Definition Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69:89–95.

Domanski D, et al. MRM-based multiplex quantitation of 67 putative cardiovascular disease biomarkers in human plasma. Proteomics. 2012;12(8):1222–43.

Eiseman E, Bloom G, Brower J, et al., editors. Case studies of existing human tissue repositories: "best practices" for a biospecimen resource for the genomic and proteomic era. Santa Monica: RAND Corporation; 2003.

Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. 2010;363:301–4.

Hewitt RE. Biobanking: the foundation of personalized medicine. Curr Opin Oncol. 2011;23(1):112–9.

Hüttenhain R, et al. Recent progress in selected reaction monitoring MS-driven plasma protein biomarker analysis. Curr Opin Chem Biol. 2009;13(5–6):518–25.

Jernberg T, et al. The Swedish Web-system for enhancement and development of evidence-based care in heart disease evaluated according to recommended therapies (SWEDEHEART). Heart. 2010;96:1617–21.

Krastins B, et al. Rapid development of sensitive, high-throughput, quantitative and highly selective mass spectrometric targeted immunoassays for clinically important proteins in human plasma and serum. Clin Biochem. 2013;46(6):399–410.

Kuhn E, et al. Developing multiplexed assays for troponin I and interleukin-33 in plasma by peptide immunoaffinity enrichment and targeted mass spectrometry. Clin Chem. 2009;55(6):1108–17.

LaBaer J. Improving international research with clinical specimens: 5 achievable objectives. J Proteome Res. 2012;11:5592–601.

Malm J, et al. Developments in biobanking workflow standardization providing sample integrity and stability. J Proteomics. 2013;95:38–45.

Marko-Varga G, Ogiwara A, Nishimura T, et al. Personalized medicine and proteomics: lessons from non-small cell lung cancer. J Proteome Res. 2007;6(8):2925–35.

Marko-Varga GA, Végvári Á, Fehniger TE. A protein shake-up. Public Serv Rev Eur Union. 2011;21:250–2.

Park A. Ten ideas changing the world right now: biobanks. Time. 2009;173(11):63.

Rezeli M, et al. Development of an MRM assay panel with application to biobank samples from patients with myocardial infarction. J Proteomics. 2013;87:16–25.

Simenon-Dubach D, Perren A. Better provenance for biobank samples. Nature. 2011;475(7357): 454–5.

Vegvari A, Marko-Varga G. Clinical protein science and bioanalytical mass spectrometry with an emphasis on lung cancer. Chem Rev. 2010;110:3278–98.

Végvári Á, et al. Identification of a novel proteoform of prostate specific antigen (SNPL132I) in clinical samples by multiple reaction monitoring. Mol Cell Proteomics. 2013;12(10):2761–73.

Yabluchanskiy A, et al. Matrix metalloproteinase-9: many shades of function in cardiovascular disease. Physiology (Bethesda). 2013;28(6):391–403.

Zolg JW, Langen H. How industry is approaching the search for new diagnostic maker and biomakers. Mol Cell Proteomics. 2004;3:345–54.

Chapter 8
Protein Microarrays: Overview, Applications and Challenges

Lucia Lourido, Paula Diez, Noelia Dasilva, Maria Gonzalez-Gonzalez, Cristina Ruiz-Romero, Francisco Blanco, Alberto Orfao, Joshua LaBaer, and Manuel Fuentes

Abstract Microarrays technology represents a new tool to address high-throughput biological studies. Various types of protein microarrays based on the application, format and content find use in basic research, drug and biomarker discovery, as well as for the characterization of proteins. In this chapter, we discuss advantages and limitations of protein microarrays, their features and recent applications. We also consider the different methods to build protein microarrays and the recent advances in cell free protein expression systems to construct *in situ* protein microarrays. Finally, we describe four types of self-assembled protein microarrays: PISA (protein array to protein Array); DAPA (DNA to Protein Array); PuCa (*in situ* puromycin array) and NAPPA (Nucleic Acids Programmable Protein Arrays) and the recent applications of this latter *in situ* protein array.

Keywords Functional proteomics • Protein microarray • Format • Content • Protein array applications • Cell-free protein synthesis • *in situ* protein microarrays

L. Lourido • C. Ruiz-Romero • F. Blanco
Instituto de Investigación BioMedica A Coruña,
As Xubias de Arriba 84, 15006 A Coruña, Spain

P. Diez • N. Dasilva • M. Gonzalez-Gonzalez • A. Orfao
Departamento de Medicina, Servicio General de Citometría,
Centro de Investigación del Cáncer, IBSAL, Universidad de Salamanca-CSIC,
Campus Miguel de Unamuno S/N, 37007 Salamanca, Spain

J. LaBaer, MD, Ph.D. (✉)
Personalized Diagnostic Laboratory, Biodesign Institute,
Arizona State University, Tempe, AZ, USA
e-mail: Joshua.labaer@asu.edu

M. Fuentes, Ph.D. (✉)
Departamento de Medicina, Servicio General de Citometría,
Centro de Investigación del Cáncer, IBSAL, Universidad de Salamanca-CSIC,
Campus Miguel de Unamuno S/N, 37007 Salamanca, Spain

Proteomics Unit, Centro de Investigación del Cáncer, IBSAL, Universidad de Salamanca-CSIC,
Campus Miguel de Unamuno s/n, 37007 Salamanca, Spain
e-mail: mfuentes@usal.es

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_8

8.1 Introduction

DNA arrays and next generation DNA sequencing technologies have found wide use in the detection of nucleic acids, revealing information about the transcriptional states of biological samples in a massive scale; however, gene expression provides only general and limited information about the function of the gene products. Moreover, nucleic acid measurements provide no information about the regulation of protein activity, which is markedly affected by posttranslational modification (PTM). A gene's function is directly manifested by the activity of its translated protein. Therefore, the detailed analysis of protein function provides a better knowledge related to the biological state of the cells (Gygi et al. 1999; Bertone and Snyder 2005).

Despite the cell biology knowledge achieved from decades of molecular biology and genetics, only a small portion of the human protein complement is understood at the biochemical level. As a reflection of the new era of research at scale, proteomics – the large scale analysis of proteins – is maturing in bringing methodology to identify, quantify and characterize the functions of all the proteins involved in biological processes (Bertone and Snyder 2005; Yu et al. 2011).

The complexity of the human proteome requires high-throughput (HT) approaches to define and study the human proteome profile. During last decade, protein microarrays have emerged as a useful tool for the analysis of the proteome at scale. Currently, protein microarrays have been successfully applied in the study of biomarkers, post-translational modification of proteins, and various types of interactions with proteins. Protein microarrays have shed light on the biological roles of proteins involved in disease (Merbl and Kirschner 2011; Hanash 2003; Dasilva et al. 2012). In this chapter, we review protein microarray technology, including the classification of protein arrays, their recent applications and challenges of this technology to address the study of human proteome.

8.2 Protein Microarrays

A classical proteomics approach involves the identification of individual proteins in a protein mixture (e.g., cell or tissue lysate) with some characterization of the quantity of each protein species. Separating the sample into fractions to simplify its complexity often precedes this type of analysis. Protein separation may be accomplished with 2D gel electrophoresis (2D-GE) or multi-dimensional liquid chromatography, and the subsequent identification of protein is done by mass spectrometry (MS) (Bertone and Snyder 2005; Gonzalez-Gonzalez et al. 2012). However, although recent advances have improved the sensitivity and the reproducibility of these techniques, they are not readily implemented in a HT format (Bertone and Snyder 2005; Gonzalez-Gonzalez et al. 2012). Moreover, the majority of the available methods to

Table 8.1 Advantages and limitations of protein microarrays

Advantages	Limitations
Protein arrays allow monitoring several proteins in the same assay (HT technology)	Protein arrays require validation experiments because of false positives can be detected
Wide range of applications: serum screening, biomarker discovery and functional proteomic studies	The highest array reported until date included only 9000 different proteins
Easy control of experimental conditions	Whole eukaryotic protein arrays still have not been reported
Low sample consumption	Difficulty to control post-transcriptional modifications
Fast	Arrayed proteins may not be functional on the surface
Very sensible comparing with other HT technologies	Lack of standard protocols

study proteins require denaturing the sample and thus functional characterization is not possible (Bertone and Snyder 2005; Gonzalez-Gonzalez et al. 2012).

In contrast with other proteomic strategies, protein microarrays avoid pre-fractionation of the sample. Thus, complex and non-fractionated proteome mixtures, such as serum, plasma, urine and tissue extracts, can be directly used for experimentation (Hanash 2003; Hanash et al. 2008) (Table 8.1). For this reason, among others, protein microarrays offer a powerful technology for functional proteomics analysis in HT format.

Microarray technologies, like DNA arrays, utilize densely-printed micro-spots of capture ligands immobilized onto a solid support that are exposed to samples containing corresponding binding molecules (often referred to as queries), allowing the simultaneous analysis of thousands of capture targets within the same assay (LaBaer and Ramachandran 2005). Roger Ekins and co-coworkers described these binding events based on miniaturization as the key parameter. They predict that a system that uses small amounts of capture molecules and a small amount of sample can be more sensitive than a system using a hundred times more material. This is true if $K<0.1$ where K is the affinity constant between ligand and target. The capture ligand is presented in a confined area of the array, reducing its diffusion. The binding event with its specific target takes place with the highest possible capture molecule concentration and therefore, the highest signal intensities and optimal signal-to noise ratios can be achieved in these small spots (Ekins et al. 1990; Ekins and Chu 1992). An immunoassay in an array format displays sensitivities in the pM to fM range, enabling testing low-abundant (pg/mL) analytes in crude proteomes with a small volume of sample. In many cases, the sample to test is minimal so protein microarrays show a relevant advantage in clinical applications.

Thus, protein array technology addresses the necessity of having a multiplex and highly sensitive protein assay capable of handling and resolving complex proteomes with limited available sample (Borrebaeck and Wingren 2009; Matarraz et al. 2011).

8.3 Array Format

Multiplex protein arrays can be prepared in two major formats in which the miniaturized assay is performed: planar arrays (such as on glass slides) and bead arrays, on which the proteins are attached to addressable beads. Here, we briefly have described some features of both formats.

8.3.1 Planar Array

Two dimensional planar multiplexed assays consist of high-density microspots of ligand (protein/peptide/aptamer/tissue) (<250 μm diameter) immobilized onto a solid support and separated by the minimal distance of 300 μm, which allow a density of >1,000 spots/cm^2 (Bertone and Snyder 2005; Matarraz et al. 2011; Ellington et al. 2010).

8.3.1.1 Array Chemistries

In planar microarrays, there are some parameters that influence the robustness of the assay performance: spot size and morphology, total ligand binding capacity, background signal, limit of detection and spot reproducibility (Dasilva et al. 2012; Ellington et al. 2010).

The greatest challenges in protein immobilization technology are the retention of natural protein folding, functionality and capacity. As binding proteins to a surface can alter these parameters, the first key step for success in a planar protein array is to define the optimal surface and protein immobilization strategy (Gonzalez-Gonzalez et al. 2012; Rusmini et al. 2007). Many materials and surface chemistries have been used for building planar arrays, ranging from PVDF membranes to glass or gold slides. Glass slides treated with organosilanes are very commonly used and they are considered suitable substrates for protein immobilization.

The most often used strategy for protein immobilization is the use of a covalent-attachment, using a wide variety of chemically activated surfaces (e.g. amine, aldehyde, etc) (Table 8.2). The abundance of lysines present on the exterior of the proteins makes the amine chemistry one of the most popular strategies to immobilize proteins.

On one side, N-hydroxysuccinimidine ester (NHS) is the most common agent to establish strong bonds with protein amine groups and its use was demonstrated by Patel et al. Aldehyde-glass slides have been also demonstrated by MacBeath and Schreiber to be a feasible strategy to immobilize large proteins like bovine serum albumin (BSA). On the other side, bioaffinity immobilization by complex avidine-biotin also offers another option to achieve a correct protein attachment. This technique permits an oriented immobilization as well as it is a reversible method, which allows repeating the use of the same surface.

Table 8.2 Available functionalities in planar arrays according to different functional groups at protein surface

Side groups	Amino acids	Surfaces
-NH_2	-Lys, hydroxi-Lys	Carboxylic acid
		Active ester
		Epoxy
		Aldehyde
-SH	Cys	Maleimide
		Pyridyl disulfide
		Vinyl sulfone
-COOH	Asp,Glu	Amine
-OH	Ser,Thr	Epoxy

In some cases, the immobilization of proteins is directed by physical interactions on hydrophobic (nitrocellulose, polystyrene) or positively charged (polylysine, aminosilane) surfaces. This is a weak bond which occurs by adsorption. As a result, proteins can be randomly oriented onto the heterogeneous surface and this fact may lead to the loss of the functional protein sites, which are inaccessible (Rusmini et al. 2007).

8.3.1.2 Array Printing

Another important variable to build planar arrays is the selection of the printing method. Briefly, the printing methods can include contact and non-contact printing. The size, morphology and reproducibility of the spots on the surface will depend on the deposition method selected.

In the contact printing, the tiny pins transfer nanoliter volumes of the printing mix on the surface, with the final transfer volume depending on the size of the pin and the length of time that pins contact the surface. Both solid and quill type pins are available for printers, however with the viscous nature of printing mixes that contain proteins, the quill type pins often do not deliver reproducible volumes. Alternatively, non-contact deposition technologies utilize capillaries or inkjet technology to deposit picoliter droplets onto the surfaces. In theory, this method decreases spot-to spot variability and the variability between batches (Ellington et al. 2010; Glokler and Angenendt 2003).

8.3.1.3 Assay Execution

There are several factors, which must to be considered before executing protein microarrays arrays to evaluate the printing reproducibility and the consequent assay.

On one hand, spotting buffer composition can influence the protein stability, the protein binding capacity to the surface and therefore, the quality of the spots produced. For this, there are many different buffers with different pH, which can be used for arraying (carbonate, PBS, citrate, acetate buffer…) samples on the surface. The choice will depend on the nature of the printed analyte (Kusnezow et al. 2003).

On the other hand the morphology of the spots will depend on sample viscosity and printing humidity. Sometimes, protein microarray production runs take a long time, depending on the amount of features on the microarray and batch size. In this sense, sample evaporation could lead to a gradient of concentration during the print run or in the worst case to blackout of print head nozzles due to salt out effects. Moreover, a higher viscosity reduces the time of sample drying. To achieve a highly reproducible microarray quality it is of prime interest to reduce this evaporation to the minimum (Gutmann et al. 2005). For this aim, the humidity along the printing must be checked according to the features of the analyte to print. In some cases, some hygroscopic additives like DMSO or glycerol can be added to prevent the sample evaporation and improve the stability of the samples (McQuain et al. 2003).

Finally, after arraying, the use of efficient blocking of reactive surface groups is critical for a reduced the unspecific binding to the surface (background). In this sense, classical blocking buffers as bovine serum albumin (BSA) at different concentrations or milk powder are very used in protein arrays to reduce the background (Kusnezow et al. 2003).

8.3.1.4 Assay Detection

The detection of interactions depends on the kind of assay performed, and may employ fluorescence, chemoluminiscence, radioisotope labelling or label-free methods, such as surface plasmon resonance or imaging atomic force microscopy (Bertone and Snyder 2005; Gonzalez-Gonzalez et al. 2012).

Fluorescence compounds are often the most useful reporter applied to detect the protein-protein interactions. The resulting signal confers high sensitivity and wide dynamic range (approximately 5 logs). Suitable fluorescence readout systems, such as high-density microarray scanners, can be used to detect when a fluorescent query is stably bound to a particular feature on the array. Signal quantification is then analyzed by specific software. Currently, the automation of these methods has increased the throughput of the planar arrays.

8.3.2 Beads Arrays

In this section, the characteristics and applications of beads arrays will be discussed as well as the differences between beads and planar arrays.

The diameter of beads typically varies from 0.02 to 0.6 μm. The beads with a diameter >0.1 μm are called microspheres and those <0.1 μm are referred to as nanoparticles (Casado-Vela et al. 2013). This size of the beads may impact the number of analytes immobilized on the surface.

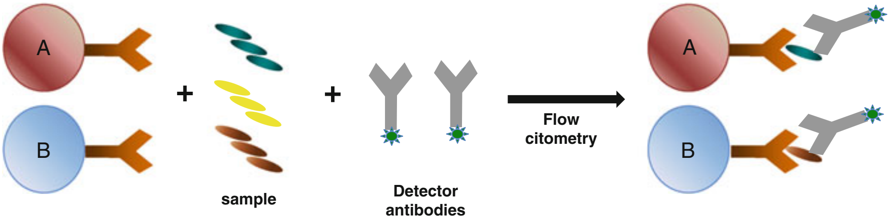


Fig. 8.1 Schematic view of suspension arrays resulting of the combination of two different fluorescent dyes (*A*, *B*). Assays are analyzed by coding attributes, and flow cytometry is used to detect assay specific fluorescence signal

Target proteins are immobilized on the bead surfaces, often using chemistries similar to planar arrays. In order to test multiple proteins simultaneously, the beads used for each unique protein must be addressable by some kind of bar-coding. Most often, this is accomplished with the use of coloured coding, using different colours and multiple intensities of each colour. Beads are filled of one or more fluorescent dyes and the surface is functionalized and coated with a capture molecule to bind in an efficient way the specific analytes in a biological sample (Casado-Vela et al. 2013; Kellar et al. 2001, 2006) (Fig. 8.1).

The immunodetection is accomplished by flow cytometry in which one or more lasers excites the internal dye(s) of the bead and a detector captures the colour profile, thus reading the "bar-code" on the bead and identifying the corresponding target protein. A second laser and detector excite and read the fluorescent dye linked to the query molecule. These captures or interaction events are assigned according to flow cytometry principles: assay-specific beads are distinguished by either light scatter or internal fluorescent ratio, and analyte-dependent signal is generated by the fluorescence generated by the capture event (Ellington et al. 2010).

This technique is based on fluorescence cell-sorting which has been used for more than 20 years in the clinical field (Casado-Vela et al. 2013). However, multiplex fluorescence bead assays were first reported in 2001 to identify and quantify cytokines in serum samples. In these naïve approaches relatively few simultaneous events were described.

Currently, uniquely bar-coded fluorescence microspheres available for Luminex Corporation allow more than 500 analytes per assay. Although the number of detected proteins in planar arrays is significantly higher than beads arrays (thousands vs. hundreds), the flexibility of the bead-based array format in some settings adds an important dimension to HT analysis.

In relation to sensitivity, the lower detection limit reported for bead fluorescence arrays is 1.2 pg/mL with a dynamic range of up to 55-fold reported by Won using fluorescence beads. This is 10-fold change more than it is reported from LC mass spectrometry analysis with a complex protein sample (Casado-Vela et al. 2013; Ellington et al. 2010). A advantage regarding to planar array is a better feasibility and accuracy of the detection since, thanks to flow cytometry, multiple independent measurements may be achieve within each microsphere population.

Therefore, beads protein microarrays combine the simplicity of immunoassay with multiplexing capability and sensitivity of protein microarrays. Almost all of the work reported thus far in the field use bead -based arrays to identify and quantify particular analytes in serum, blood or other biological fluids (Schwenk et al. 2008). These assays, like planar assays, require a minimal quantity of the biological sample to carry out the analysis. In such studies, the color-coded beads arrays are coated with antibodies with specificities against particular targets. In this assay, the biological sample is labelled with fluorescent dye by randomly chemically linking to all proteins in the sample with amino coupling chemistry. Any dye labelled protein binding to beads coated with a specific antibody will be read out as a signal attached to beads with the corresponding color code. For these assays, the accuracy and the reproducibility of the results rely on the specificity and the quality of the antibodies employed and the labeling efficiency of the proteins in the sample. A significant concern is that many antibodies exhibit cross-reactivity, which will give false readouts (Poetz et al. 2005; Schwenk et al. 2007). For this reason, many efforts must be made to demand for application-specific antibody validation (Stoevesandt and Taussig 2007).

Gevaryas et al. perform a quality control analysis between planar and bead capture protein array measuring the immune response profiles in blood samples from two large clinical studies on prostate cancer.

First, they studied the reproducibility of these protein arrays with the same samples. Then, they compared the results obtained for the same targets and samples using both platform arrays.

On one hand, their results show that both approaches quantify changes ranking 4–5-fold in the composition of the samples. On the other hand, the planar array and beads arrays that they used had good reproducibility ($R^2 = 0.77$ and 0.75 respectively) but they did not agree in around 50 % of the 57 selected proteins for this study. Therefore, they concluded that these assays highlight the need to check carefully the quality control of these assays to obtain reproducibility (Ghevaria et al. 2012).

In a recent work, Teilacker et al. described a combination platform between planar and microspheres arrays where four antigens were interrogated as a model system for multiplexed protein detection.

They used populations of fluorescent encoded microbeads conjugated with biotinylated capture antibodies and then immobilized in a flow cell.

First, the biotinylated capture antibody was conjugated to the encoded microbead and covalently coupled to flow cell. The fluidic device was consisted in carboxilate glass coverslip, a thin double-coated adhesive silicone gasket with 12 cutouts for the flow channels, and an aluminium plate with ports for tubing connection to a syringe pump.

Then four antigen solutions were spiked in serum or BSA and introduced into each channel. Finally, fluorescence imaging of the encoded microbeads was performed on an epifluorescence microscope.

They showed that the sensitivity of this method was comparable to the sensitivity obtained by enzyme-linked immunosorbent assay (pg/mL) using only 5 μL of the sample for each flow channel. With this approach, they also reduced the average area of the spots and, therefore, a subsequent reduction of sample volume and reagents.

This work leads to a great achievement towards the miniaturization because they managed to print 250 different analytes in the same area where typically, in planar arrays, a unique spot is printed (Theilacker 2011).

8.4 Content

The content of protein microarrays could have a wide diversity, from antibodies and cellular lysates to recombinant proteins. In fact, some authors classified protein arrays according to the content: Assembled arrays or self-assembled arrays. Here, it will be briefly reviewed only a few aspects because mostly of the differences with other classifications are based on nomenclature instead of methodological aspects.

8.4.1 Assembled Arrays

In these kinds of arrays, the target proteins in the array are typically antibodies, purified proteins or lysates, which are immobilized onto a functionalized surface.

8.4.1.1 Antibody Arrays

Antibody arrays are generated by printing analytes specific reagents (ASR) onto the array surface (either planar or beads; Fig. 8.2). Thus, these arrays specifically target those analytes for which there are antibodies printed on the arrays (LaBaer and Ramachandran 2005; Matarraz et al. 2011). These arrays are normally used to

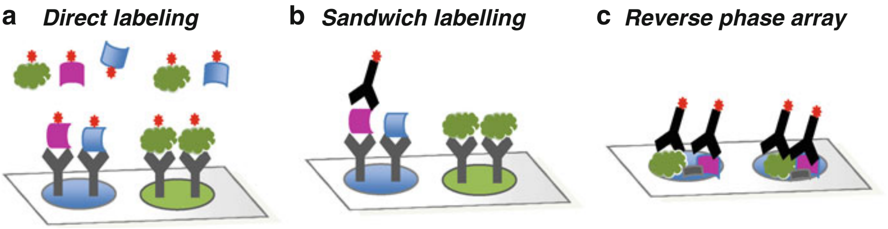


Fig. 8.2 Schematic view of different types of planar assembled arrays. (**a**) A targeted and competitive assay where proteins are directly labelled with a fluorophore; (**b**) A targeted assay with capture antibodies bind unlabeled proteins and these are detected by other labelled antibody; (**c**) Reverse phase array where protein mixtures are directly attached on the surface array and detected by other proteins or labelled antibodies

identify and quantify the presence of multiple different proteins simultaneously. They are analytical arrays whose principal application is the detection of differentially expressed proteins and their abundance in different samples. Analytical arrays are commonly used to identify biomarkers, which are biometric measurements, including molecular signatures, that predict a biological or clinical condition (e.g., healthy/pathologic), often with potential diagnostic or prognostic value (Borrebaeck and Wingren 2009).

As noted above, the accuracy of such arrays depend highly on the specificity and affinity of the antibodies. Thus, monoclonal antibodies are commonly used in this setting. However, the specificity of monoclonal antibodies can vary significantly; many will bind non-targets (**27**). Moreover, producing and qualifying monoclonal antibodies is expensive and slow, and there are many analytes for which specific monoclonal antibodies cannot be found.

Recently, there has been an increased drive towards developing analyte-specific reagents alternatives to monoclonal antibodies as recombinant antibodies. Recombinant antibodies are produced *in vitro,* in a method that does not require the use of animals, potentially providing a less expensive method to obtain antibodies. Recombinant antibodies are cloned genes encoding fragments of antibodies, which maintain the recognition capacity for the antigen. These fragments are often expressed as a single chain fragment variable (scFv) or antigen binding-fragment (Fab) (Dahan et al. 2007).

Phage display is a widely used technology to screen recombinant antibody libraries for specific molecules that recognize a desired antigen. This molecular technique takes advantage of the replication system of phages to produce different variants of the same protein or peptide of interest. The nucleotide sequence encoding the scFv or Fab is inserted into the phage genome as a fusion to a gene encoding a phage coat protein. This fusion ensures the display of recombinant antibody proteins at the surface of the mature phage. Then the antibody-displaying phage are exposed to the desired target antigen, which is immobilized on a surface (Dahan et al. 2007). This allows fractionation of the phage bearing antibodies that bind the antigen from those that do not. An advantage of this approach is that the genes encoding the successful binders can be recovered from the phage and used to produce recombinant antibodies is a wide variety of protein expression systems. However, the selection process for finding high affinity binders may require several selections cycles to achieve the needed enrichment and sometimes requires the introduction of random mutations, followed by selection to “mature” the binders to achieve acceptable specificity.

Carlsson et al. produced capture protein microarrays using recombinant scFv antibodies, and demonstrated their use for detecting changes in the levels of several interleukins and complement proteins among 40 samples from metastatic breast cancer serum and healthy donors. This same group also used recombinant antibody arrays to classify serums from systemic lupus erythematosus (SLE) and systemic sclerosis (SSc) patients (Carlsson et al. 2011).

Due to the challenges associated with producing both recombinant and monoclonal antibodies, there has been a renewed interest in polyclonal antibody reagents. Most investigators agree that polyclonal reagents, produced by the traditional method of inoculating rabbits with whole protein in adjuvant, produce reagents that have too much cross reactivity for protein microarrays, and, in any case, cannot be adequately characterized as reagents.

Recently, Larsson et al. have developed a multiplex immunization strategy for generating something that they refer to as monospecific antibodies (msAbs). First, they select the Protein Epitope Signature Tags (PrEST), which are 100–150 polypeptide sequences predicted to be unique in the proteome, the Protein. The lack of homology to other proteins and the shorter size of these PrEST is believed to contribute to more specificity in the resulting reagents (Larsson et al. 2006). Then, these fragments are cloned with N-terminal fusion, expressed in *E. Coli* and purified. Multiple purified PrESTs are mixed at equal ratios to create a multiplex antigen to immunize animals and create the polyclonal antiserum. Then, the antiserum is processed and purified over individual PrESTs to recover the PrEST-specific antibody. This strategy allows a reduction of the cost and the number of animals needed for HT antibody production. One potential limitation of the PrEST approach is that, unlike monoclonal and recombinant antibodies, the final reagent is not renewable.

To prove their strategy, in 2009, Larsson et al. selected two non-overlapping PrEST of Cytokeratin-17, as a model system to study the specificity and cross reactivity of five antibodies generated towards each PrEST.

Using planar arrays, all antibodies recognized their respective antigen but one of them showed significant binding to a third Cytokeratin-17 PrEST included on the array and overlapping amino acid sequence in both two PrEST investigated. By beads arrays, they found that these antibodies recognised same parts of the C-terminus of each PrEST.

The resulting affinity-purified antibodies were also analyzed Western blotting, immunohistochemistry and immunofluorescence.

In these assays, except for differences in staining intensity, a similar result was obtained for all antibodies in the respective PrEST-group.

These data suggest that for targets where it is difficult to find unique sequence regions it is possible to instead raise family specific antibodies recognizing a defined group of proteins rather than a specific target.

The production of affinity reagents by any method results in products falling into a wide range of affinities and specificities; therefore, it is essential to characterize each ASR with other techniques (western blot, immunoprecipitation followed by mass spectrometry, etc.) before use in protein arrays (LaBaer and Ramachandran 2005).

Assembled Array Signal Detection

Brief mention should be made of the methods for detecting the binding of analytes to captures. As mentioned above, analytes in a sample are commonly labeled directly with a fluorescent (or other) marker molecule. If one possesses more than one affinity reagent to a target of interest, and the reagents recognize different non-competing epitopes, then an indirect sandwich assay can be used to detect the analyte. Moreover, there is in increasing interest in developing detection methods that do not rely upon any labels at all.

The direct label is the simplest and most direct approach because nearly all proteins in a sample can be tested simultaneously; providing they are adequately labeled by the marker and there are corresponding capture reagents on the array. Moreover, direct labeling also allows for direct comparisons of multiple samples on the same array by labeling each sample with a distinguishable marker. For example, one might compare two time points or biological conditions by labeling the samples with different color labels. However, the specificity of detection using direct labeling relies entirely on the corresponding ASR for each analyte, which may lead to false signals in some cases.

With indirect labeling, the marker is attached to the second affinity reagent or to a secondary reagent that recognizes it (e.g., Alexa388-labeled anti-mouse IgG recognition of an analyte specific monoclonal antibody). Consequently, the indirect assay is more specific than direct labelling because it requires two ASRs to recognize the specific analyte to observe signal. However, this assay has two main drawbacks: finding a matched pair of antibodies that work well together can be very challenging, and there appears to be a practical limit of less than 40 analytes that can be measured simultaneously using antibody pairs. Interference and cross-reactivity become an increasing problem as the number of antibodies added to a common detection mix increases. Huang et al. used a sandwich assay to measure the levels of 24 cytokines in two biological conditions. There is a vast collection of antibodies that recognize cytokines with well-established target epitopes, but for most other antigens, it is difficult to find compatible antibody pairs to carry out a sandwich assay (Gonzalez-Gonzalez et al. 2012; Huang et al. 2001).

Antibody microarrays are well suited as screening tools for discovering disease-specific biomarkers owing to their potential to measure thousands of proteins in rapid, low volume assays. A number of reports have applied protein microarrays to biomarker discovery in cancer.

Recently, Gao et al. measured the abundance of eighty-four proteins in a two-color assay to compare chronic obstructive pulmonary disease (COPD), newly diagnosed subjects with lung cancer and healthy controls serums (Gao et al. 2005) being each sample compared to a pooled reference sample (consisting of a mixture of all of the sera).

The values determined were the normalized average of base-2 logarithms of the intensity arising from the individual sample divided by the intensity arising from the pooled sample, which was measured as Cy3 and Cy5 fluorescence, respectively. With this approach, they found that using an analysis variance model, 7 antibodies showed significant differences between both lung tumor patients vs. normal controls and lung tumor patients vs. COPD patients (Gao et al. 2005).

Wittekind & colleagues also reported a study where proteins from 30 normal and hepatocellular liver were differently labelled with Cy3 and Cy5 and hybridized on a nitrocellulose protein microarray made up 83 different antibodies.

Proteins of each condition (1 mg/mL) were labeled with NHS-ester activated Cy3 or Cy5 and mixture of equal concentrations to incubate on the array. The ratios between dyes were determinated for the individual proteins. To determine which proteins were found to be differentially expressed, a cutoff level was fixed using a hierarchical model.

Using a stringent interval of 0.4–1.9 corresponding to 2.5 SDs, five proteins were classified as differentially up- regulated: IGFII, ADAM9, STAT3, SOCS3, and cyclin D1 and four proteins (collagen I, SMAD4, FHIT, and SOCS1) as differentially down-regulated. These data were confirmed using western blot analysis of selected proteins using identical antibodies.

To test the sensibility of the array, purified proteins were spotted demonstrated high sensitivity and specificity of the protein microarray system, with a detection limit of 6.25 pg of spotted proteins.

To test the specificity of the microarray data, a fluorescent dye reversal experiment was performed. HCC and normal tissue were labelled with Cy3 and Cy5, respectively. Similar profiles of up- or down-regulated proteins were obtained, irrespective of the dye used (Tannapfel et al. 2003).

Amonkar et al. described the development and preliminary evaluation of a multianalyte profile that can classify women suspected of having ovarian cancer, into those with and without ovarian cancer.

In this work, sera from 176 cases representing all stages (I,II,III,IV) of epithelial ovarian cancer and 187 controls from women presenting, the most common benign ovarian conditions. These samples were assayed in a protein array consisted in a panel of 104 antigens, 44 autoimmune and 56 infectious disease markers

Analytes were quantified by reference to 8-point calibration curves and machine performance was verified using three quality control (QC) samples for each analyte. Almost all the samples were analyzed in two rounds and the QC samples generally had coefficients of variance below 15 %.

Then, they built many analyte classifiers and selected the best model using bootstrap performance. Finally, using a testing set of 245 samples, they built an 11-analyte classifier had 91.3 % sensitivity and 88.5 % specificity.

Sreekumar et al. used antibody arrays made up 146 distinct antibodies to monitor changes in the levels of proteins in colon carcinoma cells after quimiotherapy (Sreekumar et al. 2001).

In this work, control and stimulated cells were labeled separately using either Cy5 or Cy3 dyes and incubated on the array. Cy3:Cy5 ratios were determined for the individual proteins and a cutoff of 1.15, value was used as a criterion to define proteins considered differentially expressed and a *P* for each of the differentially regulated proteins was calculated using an unpaired *t* test

The validation of protein microarray data was done by fluorescent dye-reversal and immunoblot analysis to the selected proteins.

Belov et al. developed an nitrocellulose antibody microarray to immunophenotype leukaemia and lymphomas according to the abundance of a panel of 60 antigens or cluster of differentiation (CD) characterized at the surfaces of lymphocytes (Belov et al. 2003).

They estimated the average number of cells bound *per* dot (determined microscopically), which correlated well with average binding density values from the image analysis software, and generally correlates with results from flow cytometry.

They concluded that the CD antibody microarray enables rapid, concurrent screening of leukocyte suspensions for expression of many CD antigens. And that, in contrast to flow cytometry, with this platform, cells captured on the CD antibody microarray can be imaged directly with an optical scanner without staining or labeling, and image analysis software gives immediate results.

In addition to protein based affinity reagents, there is an increasing interest in the use of nucleic acid-based affinity reagents. The most common of these are referred to as aptamers and they are constructed from single stranded DNA or RNA, often with modified nucleotides, which fold into conformations that bind specifically to antigens. They are chemically synthesized and in vitro selected by cycles of binding amplification with a SELEX technique (reviewed in (Ellington et al. 2010)).

In 2010, Gold et al. created a new class of aptamers with modified nucleotides, the Slow Off-rate Modified Aptamer (SOMAmer). With the incorporation of these quemically-modified nucleotides (SOMAmers) into SELEX experiments, they measured 813 proteins with low limits of detection (1 pM median), 7 logs of overall dynamic range (100 fM–1 mM), and 5 % median coefficient of variation.

Briefly, in SOMAMER technologie, the sample is incubated with a mixture of SOMAmers each containing a biotin, a photocleavable group, and a fluorescent tag at 5′-end followed by capture of all SOMAmer-protein complexes on streptavidin beads. After stringent washing of the beads to remove unbound proteins and labeling of bead-associated proteins with biotin under controlled conditions, the complexes are released from the beads back into solution by UV light irradiation and diluted into a high concentration of dextran sulfate, an anionic competitor. The biotin that was originally part of the SOMAmer remains on beads. The anionic competitor coupled with dilution selectively disrupts non-cognate complexes and because only the proteins now contain biotin, the complexes are re-captured on a second set of beads from which unbound SOMAmers are removed by a second stringent washing. The SOMAmers that remain attached to beads are eluted under high pH-denaturing conditions and hybridized to sequence-specific complementary probes printed on a standard DNA microarray.

In this work, they demonstrated the specificity of SOMAmers for the proteins they were selected against but they need to validate and standardize SOMAmer-based measurements and expand these studies to understand the specificity of SOMAmers for close homologues and alternate forms, such as the products of alternative splicing, post-translational modifications, and proteolytic cleavage.

With this method, De Groote et al. identified several proteins that exhibit significant expression differences during the intensive phase of tuberculosis therapy. However, these findings require future testing in properly designed validation studies using independent sample test sets with proper disease controls.

8.4.2 Reverse-Phase Arrays

The concept of ‘Reverse-phase’ protein microArray (RPA) assays is essentially the inverse of the capture arrays. Instead of testing each sample (e.g., clinical sample) for many possible analytes, each RPA tests many samples for the abundance of a specific analyte. Essentially, it is a sophisticated micro-scale version of a dot blot. Complex protein mixtures (such as cellular or tissue lysates, or biological fluids such as serum, etc.) are printed to a substrate in a defined array pattern and then probed with antibodies or other affinity reagents that are highly specific for analytes of interest. The most critical aspect of RPAs is the extensive validation of the antibodies to ensure that they do not have any cross-reactivity with other proteins that may be present in the lysate. RPAs can generate 1,000 times more data using 10,000 times less sample volume than an ordinary western blot (Fig. 8.2). Large-scale sample collection is a labor-intensive and time-consuming process; however, the information yielded from RPA assays, enables researchers to evaluate theoretical protein pathways experimentally in a HT format (Gonzalez-Gonzalez et al. 2012; Spurrier et al. 2008).

One of the limitations for this technique is the dynamic range because the ability to detect low abundance proteins in complex mixtures is a challenge for this kind of array (Dasilva et al. 2012). Several techniques have been developed for the detection of these low abundance proteins. As in the analytical arrays, fluorescence is commonly used for standard signal readout. During analysis, signal intensity among spots is often normalized with the total protein content per spot which can be measured with different dyes (Sypro Ruby, colloidal gold, etc). The dye choice to quantify the protein content will depend on type of sample being investigated, the sensitivity of the stain, the material of the surface and of the instrument detectors (Gallagher et al. 2011).

Paweletz et al. immobilized protein lysates from microdissected histologically normal prostate epithelium, prostate intra-epithelium neoplasia (PIN) progression and invasive neoplasia. Using a Wilcoxon test for the statistical analysis, they observed a significantly increased (p value < 0.03) in the phosphorylation of AKT protein which was associated with cancer progression in the epithelium transition along the disease as well as a decreased phosphorylation of ERK (p.value < 0.01) for the three comparisons. (Paweletz et al. 2001). They also validated these results by Western Blot assay.

Cid et al. were the first to use RPA to study pathogen-host protein interactions. In this work, they studied the presence of posttranscriptional modifications in effector proteins, T3SS proteins, from different mutants of *Salmonella typhimurium* when they infected *in vitro* HeLa to understand signaling events that take place along *Salmonella* infection and the intracellular survival of this bacteria into the cells.

Lysate collection representing all infection conditions were printed and using several validated antibodies against phospo-epitopes, they show a comparative results among the different assays according to abundance proteins or posttranscriptional modification (Molero et al. 2009).

8.4.3 Self-Assembled Protein Microarrays

These arrays focus mainly on identifying and characterizing the specific function of proteins, as well as their interactions with other molecules (including proteins, peptides, small molecules/drugs, enzyme-substrates or nucleic acids) (LaBaer and Ramachandran 2005). These functional protein arrays also allow the detection and identification of post-translational modifications (PTMs), such as glycosylation, phosphorylation and acetylation, which typically modulate the protein's function, regulation and/or turnover (Casado-Vela et al. 2013).

The first critical step to build protein microarrays is to display proteins on a solid surface for the detection of their biochemical activities in a multiplex manner. Notably, the intrinsic properties of proteins, particularly their highly variable biochemical properties, make building protein arrays more challenging than building nucleic acid arrays, which have very consistent chemical properties (Bertone and Snyder 2005; Templin et al. 2002; Rusmini et al. 2007). Briefly, some of protein properties which must be accommodated when building protein arrays include: (i) Wide variety of chemistries, affinities and specificities; (ii) different oligomerization state from monomer to multimers; (iii) different PTMs; (iv) varied protein stability, which is frequently altered when the protein is deposited or immobilized onto a surface; (v) protein production and purification in high-throughput manner with high yield could be also challenging (Gonzalez-Gonzalez et al. 2012; LaBaer and Ramachandran 2005).

The cell-based expression system and the purification to generate large quantities of proteins is usually a very tedious task and do not guarantee the functional integrity of the protein. This issue represents a bottleneck in the HT functional proteomic studies. Nowadays, it is possible achieve these drawbacks building arrays of full-length, functional proteins from a library of clones expressed *in situ*.

In the *self-assembled* protein microarrays, the protein are synthesized from their corresponding messenger ribonucleic acid (mRNA) or complementary deoxyribonucleic acid (DNA) directly on the surface of the array and the immobilization of the nascent protein is coupled in the same step in a fast manner.

On the research side, self-assembled arrays offer the detection of multiple protein interactions with low reagent consumption in a fast and low cost fashion. On the translational side, the discovery of these interactions will foster the development of new pharmaceutical targets, diagnostics and therapeutics. Thus, this technology is an attractive point of sight for the pharmaceutical industry (Dasilva et al. 2012; LaBaer and Ramachandran 2005).

8.4.3.1 In Situ Protein Expression Systems

From the discovery of in situ protein expression systems forty years ago by Nirenberg and Matthai, they have been broadly utilized in the scientific community to solve the issues of in vivo protein production. A main advantage that these systems have over in vivo protein synthesis is that the environmental conditions can be adjusted easily. Strategies to improve protein folding and posttranslational processing include the addition of a variety of reagents and folding catalysts to the reaction. But, above all, the main goal for cell-free translation systems is to synthesize biologically active proteins (Casado-Vela et al. 2013).

These in vitro expression systems exploit the ability to translate proteins using properly prepared lysates from a number of different organisms (both prokaryotic and eukaryotic), which provide the ribosomal machinery, accessory enzymes, tRNAs, amino acids and an appropriate energy source. Moreover, these lysates can be coupled with specialized RNA polymerases and the appropriate nucleotides to allow simultaneous transcription, thereby allowing full transcription and translation of proteins from exogenously added cDNA templates.

Prokaryotic cell-free expression systems can produce up to mg quantities of protein (Hunt 2005; Kigawa et al. 1999; Mijakovic and Macek 2012; Plotkin and Kudla 2011).

They are reasonably tolerant to additives (cofactors, protease inhibitors or energy sources) (Casado-Vela et al. 2013). However, some of the same limitations that bacteria have producing eukaryotic proteins, such as marked decrease in success for proteins >65 kDa, also plague cell free systems from bacteria (Casado-Vela et al. 2013). Typically, bacterial cell-free systems do not produce posttranslational modifications (PTMs), which can be either useful or not depending on the application. It is worth noting the recent introduction of highly characterized cell free systems from bacteria. These systems are produced entirely from purified recombinant proteins and ribosomal RNA. In this manner, they are highly characterized and some applications might benefit from using a system where every component is known and there is no risk of contaminants from a crude lysate (Ref).

Cell-free eukaryotic expression systems include wheat germ, insect, rabbit reticulocyte and human lysates.

Although rabbit reticulocyte produces less protein than other expression systems (0.2 μg/10 μL reaction) and is very expensive, it is commonly used for functional proteomic studies because it is very fast (2 h approx. in protein production), it has a very high success rate for most mammalian proteins, including large and membrane proteins, and it does support limited PTMs (Casado-Vela et al. 2013). Rabbit reticulocyte lysate also contains most chaperone proteins, so the there is a high likelihood that translated proteins will fold naturally and even display activity if they can act monomerically (Casado-Vela et al. 2013).

As a result of the source and method of production, each batch of rabbit reticulocyte lysate derives from a single animal and is by necessity a non-renewable resource. An additional disadvantage of this product is that there is a high degree of variability from one batch of lysate to the next, both in the yield of protein produced and sometimes in the presence of other factors that can affect downstream applications. Laboratories might have to test a dozen different batches to find one with the right activity profile needed for their experiments.

The recent development of cell free lysates from human cells addresses some of these concerns (Casado-Vela et al. 2013). These lysates are produced from a cultured human cell line, which allows for much greater control of growth conditions and significantly minimizes the batch-to-batch variation. Avoiding the presence of animal serum also reduces the likelihood of contaminants that can interfere with some assays. The yields of protein for these lysates are quite favorable and in the case of human protein production, the use of human ribosomes and chaperone proteins gives the greatest possible chance for natural protein folding (Casado-Vela et al. 2013).

8.5 Types of *In Situ* Protein Arrays

8.5.1 PISA

PISA (protein *in situ* array) was introduced by He and Taussing in 2001 and it was the first well known cell free based *in situ* protein array. This method uses PCR-amplified DNA as template. The use of PCR product obviates the need to clone the open reading frame into a plasmid, but for large-scale array production, the repetitive use of PCR to produce the template for making arrays can become both costly and lead to losses in fidelity. The DNA encoding the protein of interest contains a T7 promoter or another strong transcriptional promoter and an in-frame N- or C-terminal tag sequence for protein capture onto the surface. The tags are typically short peptides so that their sequences can be incorporated into the PCR primers. In order to perform PISA; the wells of a microtiter plate are pre-coated with a tag-capturing agent. After transcription and translation, the expressed proteins bind onto the surface through the specific tag. In the first PISA, He and Tau used DNA templates to express human anti-progesterone antibody and luciferase with 6X histidine tags. These were captured into a microtiter plate with 24 wells coated with nickel nitrilo-triacetic acid (Ni-NTA) and Ni-NTA-coated magnetic beads respectively. They checked that small quantities of these proteins can be expressed and immobilized onto the surface and that they conserved their functional features. Also, they confirmed its HT applications, such as the generation of protein arrays for non-available cloned genes or for proteins without functionally production in heterologous expression systems. They suggested the combination with *in vitro* display methodologies for HT identification. PISA opened the door to cell free production of protein arrays. It demonstrated that multiple proteins could be produced without the need to use cells for expression followed by lysis and purification to make the proteins (Gonzalez-Gonzalez et al. 2012). In 2006, Angenendt reported a PISA where 384 different proteins could be expressed from very small quantities of template (Angenendt et al. 2006; Casado-Vela et al. 2013) (Fig. 8.3).

8.5.2 DAPA

This innovative technique also was developed by in He et al. in 2008. DAPA (DNA Array to Protein array) is a technique derived from PISA but it allows for the repeated use of the same DNA template slide for printing up to 20 copies of the same protein array and DNA could be reused after prolonged periods of time. DAPA starts by spotting the PCR amplified DNA fragments encoding the tagged protein on one slide. This slide is sandwiched with another Ni-NTA slide where a tag-capturing agent immobilizes the expressed protein. A permeable membrane with the cell-free

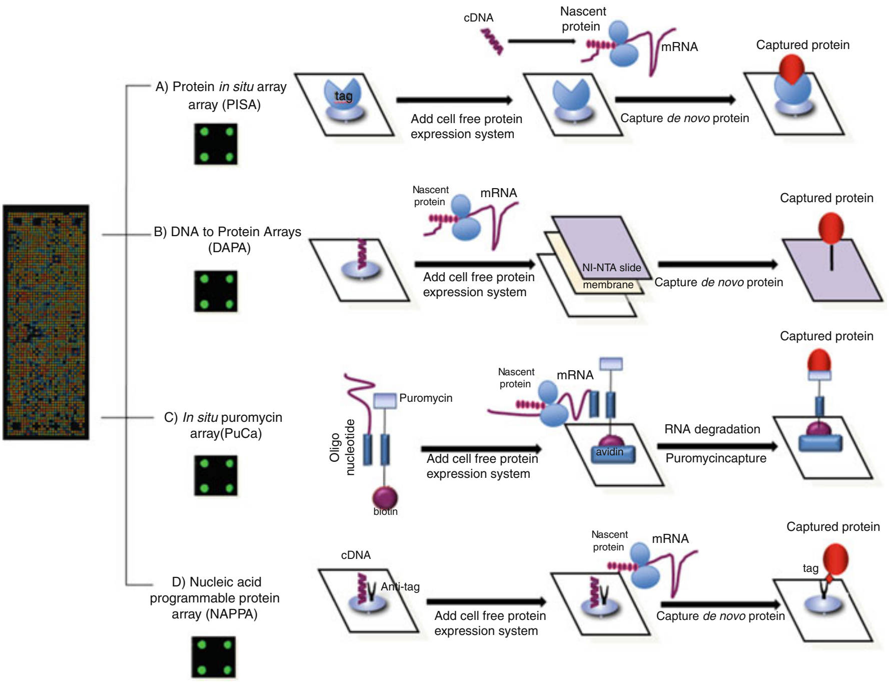

A) Protein *in situ* array array (PISA)
tag
cDNA
Nascent protein
mRNA
Add cell free protein expression system
Capture *de novo* protein
Captured protein
B) DNA to Protein Arrays (DAPA)
Nascent protein
mRNA
Add cell free protein expression system
NI-NTA slide
membrane
Capture *de novo* protein
Captured protein
C) *In situ* puromycin array(PuCa)
Oligo nucleotide
Puromycin
biotin
Add cell free protein expression system
Nascent protein
mRNA
avidin
RNA degradation
Puromycincapture
Captured protein
D) Nucleic acid programmable protein array (NAPPA)
cDNA
Anti-tag
Add cell free protein expression system
Nascent protein
mRNA
Capture *de novo* protein
Captured protein
tag

Fig. 8.3 Schematic view of four methods coupling cell-free protein synthesis to protein binding on the surface arrays. (*A*) protein *in situ* arrays(PISA); (*B*) DNA to Protein Array(DAPA); (*C*) puromycin capture protein array (PuCa) which uses mRNA as template; (*D*) nucleic acid programmable protein arrays (NAPPA). Figure adapted from Casado-Vela et al. 2013

lysate, which allows coupled transcription and translation is placed between the two slides. Proteins synthesized from immobilized DNA spots diffuse through the membrane and are bound to the surface of the capture ligand on the other slide. Even though it is used, the DAPA requires long time to express proteins and this technique is limited by the possibilities of protein diffusion during membrane penetration, especially regarding larger multimeric proteins (Gonzalez-Gonzalez et al. 2012; He et al. 2008).

8.5.3 PuCA

Puromycin capture protein arrays (PuCA) are cell-free expression protein arrays based on the affinity of puromycin by just-in time expressed peptide/protein. Tao and Zhu developed it in 2006. With this method DNA is transcribed into mRNA *in vitro*. The mRNA 3′ end is attached with single stranded oligonucleotides (ssDNA) which is complementary with another ssDNA bearing biotin and puromycin and this complex together is layered onto the chip surface. Biotin serves to immobilize the mRNA onto the coated streptavidin surface and puromycin serves as an anchor to bind the nascent protein translated when cell-free expression system is added to the array (Gonzalez-Gonzalez et al. 2012; Tao and Zhu 2006).

8.5.4 NAPPA

Although useful in research, these mentioned strategies have only been tested with relatively small numbers of proteins compared with printing purified proteins and have yet to demonstrate the robust ability to produce the high content needed to justify protein microarrays as a routine proteomics tool (LaBaer and Ramachandran 2005).

In 2004, LaBaer's lab developed a high-density self-assembled protein microarray called nucleic acid programmable protein array (NAPPA) (Ramachandran, Science). It is based on cDNA templates cloned into expression plasmids, typically using the Gateway technology, which add a transcriptional promoter and also adds an in-frame polypeptide capture tag. The requirement to clone the cDNAs into a specialized vector requires a much greater upfront investment compared with PCR. However, there are several advantages: (1) once the clone is produced as a glycerol stock it becomes a indefinitely renewable resource that can be shared with other labs; (2) if the clone is carefully sequence verified, then the resource will have long term sequence fidelity; (3) the use of plasmids removes some of the length constraints on the epitope tags, so that functional protein tags can be used. In most applications of NAPPA the proteins are fused with glutatione-S-transferase (GST); however, other tags such as flag, HA, c-myc, and Halo tag have been used in specific applications. High quality supercoiled plasmid DNA is purified from bacteria cultures and printed onto an activated ester surface along with a homo-bifunctional crosslinker, bovine serum albumin (BSA) and anti-GST antibody. BSA efficiently increased the DNA binding and reduces the unspecific interactions and anti-GST attaches the protein expressed (Ramachandran, Science). When cell-free expression system is added to the array, a coupled transcription/translation reaction is produced and the nascent protein is linked to the capture agent tag the C-terminal end assuring the complete translation of the protein (Ramachandran et al. 2008a).

In an updated method for NAPPA, LaBaer and colleagues built an array of 1,000 human genes available through the DNASU repository (http://DNASU.org) and demonstrated that 96 % of the genes showed detectable protein signal, including both soluble and membrane proteins (Ramachandran Nature Methods). They concluded that the protein size had only a modest effect with 98 % of proteins <50 kDa showing good display levels, whereas proteins >100 KDa showing success around 88 %. With this report they concluded that this method enables various experimental approaches to study protein function in HT (Ramachandran et al. 2008a).

8.6 Applications of NAPPA Arrays

Of the various *in situ* methods for producing protein microarrays, NAPPA is the only one that has been extensively used in biological and biomedical discovery experiments. To date, more than 30,000 different proteins have been produced on NAPPA arrays, including the whole proteomes of several microorganisms and 10,000 different full-length human proteins. Array production has largely been automated, so that thousands of NAPPAs can be produced per year.

8.6.1 NAPPA Arrays for Vaccine Development

In 2009 Thanawastien et al. published a report where they used cell-free expression system to develop a new HT approach called Expressed Protein Screen for Immune Activators (EPSIA) in order to identify novel bacterial immunostimulatory proteins from *Vibrio cholerae.*

Firstly, they expressed *in vitro V. cholerae* proteins from 7 ORF expression plasmid libraries. The synthesized proteins were then added to treat cultured RAW264.7 murine macrophage cells and primary peritoneal macrophage cells and culture supernatants were collected to assay the production of several pro-inflammatory cytokines by ELISA.

They found that phosphatidylserine decarboxylase (PSD) was a conserved bacterial protein capable of activating host innate immunity inducing the secretion of TNFα and IL-6, two strong pro-inflammatory cytokines.

With this approach, they concluded that EPSIA provides an approach to screening the entire protein repertoire of an infectious organism for agonists of immunological responses that can be assayed using appropriate eukaryotic reporter cell lines (Thanawastien et al. 2009).

In a most recent study, Montor et al. describe a work using NAPPA arrays to test candidate membrane antigens in *P. aeruginosa. P. aeruginosa* is a gram-negative bacterium ubiquitous in the environment which rarely causes respiratory tract infections in healthy individuals but causes life-threatening lung infections in cystic fibrosis (CF) patients. The goal of this study was to map the immune responses of patients infected with *P aeruginosa* to determine which bacterial proteins induced a strong immune response. As the focus of this study was on outer membrane proteins, which are notoriously difficult to express and purify, the use of NAPPA, which routinely expresses and displays membrane proteins, was particularly fortuitous.

For this approach, 262 from *P. aeruginosa* PAO1 ORFs encoding all of the known and predicted outer membrane proteins were successfully transferred to the in vitro expression vector pANT7-cGST and printed onto NAPPA slides. The slides were screened with independently prepared serum samples from 22 CF patients with documented *P. aeruginosa* infections and 16 non-CF with various acute *P. aeruginosa* infections as well as 15 healthy controls.

This work identified 12 proteins that triggered an adaptive immune response in a majority of the infected patients, yielding valuable information about which bacterial proteins are recognized by the immune system during the natural course of infection (Montor et al. 2009).

8.6.2 NAPPA for Protein-Protein Interaction

In 2008, LaBaer and colleagues confirmed with NAPPA arrays that protein function was maintained for printed proteins on high-density arrays. In this report they printed an array expressing 647 unique genes in duplicate and tested for several well-characterized interacting pairs including Jun-Fos and p53-MDM2. At the same time, they expressed the corresponding protein printed on the array and co-expressed the query protein by adding the appropriate cDNA to the cell-free expression lysate. Using specific antibodies against Jun, Fos and MDM2 as queries, they detected specific binding of these proteins interacting with their partners. Given that there are no simple tests to confirm protein folding, the function must be tested on a protein-by-protein basis (Ramachandran et al. 2008a). The protein function can be compromised by lacking of PTMs and by misfolding of certain domains. This folding often relies on the presence of chaperones and cofactors. The IVTT of rabbit reticulocytes used by LaBaer in all these experiments is an open system which allows the use of chaperones which may encourage folding and it is possible to add modifying enzymes or extracts, such as kinases or canine pancreatic microsomal membranes, to test the effect of post-translational modifications (Ramachandran et al. 2008a).

In 2012, Fuentes et al. published a work where they applied NAPPA to study protein-protein interactions. They extracted and purified mRNA from 450 *O. moubata* tick salivary glands. Then, they synthesized a library of cDNA transfecting the polyA + mRNA to a donor vector (pDONR222). Finally, this library was transfected into a library destination expression vector (pANT_GST), which allows *in situ* expression of GST-tagged proteins in cell-free systems. They randomly chose 480 clones which sequence had been previously validated and including these clones to build a NAPPA array. After analyzing the correct expression of the recombinant fused GST tag protein, the correct expression of tick proteins was also checked by incubating the arrays with a serum, which recognized Om44, a salivary protein from *O. moubata*. This protein is a P-selectin whose neutralization induced antibodies blocks tick feeding. To test the functionality of the proteins in the array, they performed protein-protein interaction studies with the recombinant P-selectin/Fc chimera. With this aim, the proteins on the array and P-selectin/Fc chimera were expressed *in situ* normally and in the presence of canine pancreatic microsome membranes (CMMs) and then probed with the P-selectin/Fc chimera. They found that P-selectin/Fc chimera interacted with phospholipase A2 (PLA2) expressed *in situ* on the array. This finding suggests that this secreted *O. moubata* phospholipase A2 (sPLA2) could be a potential P-selectin interacting partner (Manzano-Roman et al. 2012).

8.6.3 NAPPA Arrays for Detecting Autoimmune Response

In addition to producing antibodies against foreign molecules, the humoral immune system generates antibodies to self-proteins (“auto-antibodies”) in response to many pathological processes. It is believed that autoantibodies are generated through antigen over-expression, mutation, altered post-transcriptional modification of altered degradation released from damaged tissues which leads to their recognition by the immune system (Ramachandran et al. 2008b). The production of these autoantibodies is not common healthy individuals, so their detection might be useful as biomarkers for the presence of diseases like diabetes and cancer. Autoantibodies have several benefits which make them good biomarkers: (1) they have been detected before clinical symptoms appear (ref); (2) they are easy to detect even at low levels once their target antigen is known; (3) they are easy to collect from blood; and (4) they could be present in higher levels and with a longer half-life than their target antigens, which may only be present transiently in blood.

The first time that NAPPA arrays were used for serological screening was in 2007 by Anderson and colleagues. They studied the presence of antibodies against tumor antigens in breast cancer. p53 is a well-studied tumor suppressor in many cancers and the presence of antibodies against p53 is thought to be due to mutations in its gene which lead to alterations in its half-life. They first expressed p53 along with other three negative control antigens (S100A7,p21 and ML-IAP) in NAPPA arrays and tested it with positive and negative p53 sera. They confirmed the expression for all the proteins printed and checked the detection of antibodies against p53 and not for the antigen controls and these results were validated by recombinant p53 ELISA. In addition, they showed that p53-specific antibody levels were significantly lower in healthy donors than in breast cancer patients and the response to p53 antigen was detected in Stage II disease. They also tested the antigen sites of p53 with several antibodies which recognized different epitopes of the protein to confirm that many regions of the protein expressed in NAPPA were accessible to antibodies in serum directed to them (Anderson et al. 2008). To extend the study to autoantibody biomarker detection, they built a high density NAPPA array printing 1,117 cancer related genes of which 539 were implicated on breast cancer and tested this against melanoma, ovarian and breast cancer sera (Anderson et al. 2008).

In a later work, the same laboratories did a more extensive screen for novel autoantibodies in breast cancer. They arrayed and expressed 4988 candidate antigens to detect their autoantibodies in serum samples from breast cancer patients with stage I-III disease. This was done in a three stage design that entailed comparing cases and controls and eliminating uninformative antigens at each stage. At the final stage, slightly more than 100 antigens were tested and 28 autoantibodies were identified that distinguished benign breast disease from invasive cancer under blinded conditions (Anderson et al. 2011a).

More recently, LaBaer et al. developed a pilot NAPPA to assess autoantibodies present in juvenile idiopathic arthritis (JIA) which is a disease characterized by chronic joint inflammation in children (Gibson et al. 2012).

Related to type 1diabetes (T1D), LaBaer's lab has profiled serological autoantibodies (AAbs) from the disease by using the NAPPA strategy. A two-stage method was performed for the screening followed by a validation study. In the first stage, more than 6000 unique proteins were printed. The incubation with 50 sera from T1D patients and 20 from controls allowed the elimination of uninformative antigens. In the second stage, 750 genes were printed in duplicate. 26 proteins were identified as novel AAbs (TBCA, CDK4, CDK6, TBRG4, among others) by applying the Wilcoxon Rank-Sum Test ($p < 0.005$) to the normalized signal intensities (Miersch et al. 2013).

In 2009, Anderson et al. published a report where they developed a programmable multiplexed immunoassay for the rapid monitoring of humoral immunity, adapting the NAPPA approach to the Luminex suspension bead array platform. To accomplish this, they expressed *in vitro* proteins tagged GST or FLAG from ORFs libraries and captured them onto Luminex beads coupled with anti-tag antibodies. In order to test viral EBNA-1 and auto-antigen p53, human sera were added to the mix and the immunodetection was revealed by anti-IgG human antibody. After this study, they concluded that detection of antibodies against EBNA-1 antigen and p53 in human sera is highly reproducible and the specificity and limits of detection of the bead ELISAs are comparable to both standard protein ELISAs. Therefore, this method allows for rapid conversion of ORFeome-derived cDNAs to a multiplexed bead ELISA to detect antibody immunity to both infectious and tumor antigens (Wong et al. 2009). In 2011, Luminex suspension bead arrays were also employed by Anderson et al. to detect and quantify antibodies against several oncogenes related to human papillomavirus (HPV) 16 and associated with oropharingeal carcinomas (OPC). 40 sera obtained from OPC patients, 11 HPV16+ OPC patients and 30 healthy donors were tested. Each HPV oncogen was expressed as GST-fusion proteins and linked to anti-GST coated beads arrays. Then, the beads were polled and aliquoted to a 96-well plate. Beads were blocked, incubated with the sera diluted at 1:80 and the autoantibodies were detected with anti-IgG human-PE.

The results showed that HPV16+ OPC have detectable Abs to E1, E2, and E7 oncogenes. In a validation cohort, these proteins are significatively increased in HPV16+OPC ($p < 0.01$) compared to healthy and OPC serums. They concluded that these antibodies might be potential biomarkers for HPV-associated OPC (Anderson et al. 2008, 2011b).

8.7 Conclusions and Future Directions

Here, we have briefly reviewed protein microarray field from two major perspectives: *i.*-Key technological aspects, *ii.*-Biological applications. However, as described previously, despite the important advances in protein microarrays allowing characterization of whole human proteome is still remaining as a challenge. Then, the information provided by protein arrays about the binary interaction occurring on human proteins will provide light on the function of proteins and genes whose functions are currently unknown.

Overall, protein arrays may provide relevant information about the biological function of gene products. Although, it is still necessary to develop and optimized some key aspects on protein microarray; in addition, other proteomics approaches could provide complementary results.

References

Anderson KS, Ramachandran N, Wong J, et al. Application of protein microarrays for multiplexed detection of antibodies to tumor antigens in breast cancer. J Proteome Res. 2008;7:1490–9. doi:10.1021/pr700804c.

Anderson KS, Sibani S, Wallstrom G, et al. Protein microarray signature of autoantibody biomarkers for the early detection of breast cancer. J Proteome Res. 2011a;10:85–96. doi:10.1021/pr100686b.

Anderson KS, Wong J, D'Souza G, Riemer AB, Lorch J, Haddad R, Pai SI, Longtine J, McClean M, LaBaer J, Kelsey KT, Posner M. Serum antibodies to the HPV16 proteome as biomarkers for head and neck cancer. Br J Cancer. 2011b;104(12):1896–905.

Angenendt P, Kreutzberger J, Glokler J, Hoheisel JD. Generation of high density protein microarrays by cell-free in situ expression of unpurified PCR products. Mol Cell Proteomics. 2006;5:1658–66. T600024-MCP200 [pii]. doi:10.1074/mcp.T600024-MCP200.

Belov L, Huang P, Barber N, et al. Identification of repertoires of surface antigens on leukemias using an antibody microarray. Proteomics. 2003;3:2147–54. doi:10.1002/pmic.200300599.

Bertone P, Snyder M. Advances in functional protein microarray technology. FEBS J. 2005;272:5400–11. doi:10.1111/j.1742-4658.2005.04970.x.

Borrebaeck CAK, Wingren C. Design of high-density antibody microarrays for disease proteomics: key technological issues. J Proteomics. 2009;72:928–35.

Carlsson A, Wuttge DM, Ingvarsson J, et al. Serum protein profiling of systemic lupus erythematosus and systemic sclerosis using recombinant antibody microarrays. Mol Cell Proteomics. 2011;10:M110.005033.

Casado-Vela J, González-González M, Matarraz S, et al. Protein arrays: recent achievements and their application to study the human proteome. Curr Prot. 2013;10:83–97.

Dahan S, Chevet E, Liu JF, Dominguez M. Antibody-based proteomics: from bench to bedside. Proteomics Clin Appl. 2007;1:922–33. doi:10.1002/prca.200700153.

Dasilva N, Diez P, Matarraz S, et al. Biomarker discovery by novel sensors based on nanoproteomics approaches. Sensors (Basel). 2012;12:2284–308. doi:10.3390/s120202284.

Ekins R, Chu F. Multianalyte microspot immunoassay. The microanalytical "compact disk" of the future. Ann Biol Clin. 1992;50:337–53.

Ekins R, Chu F, Biggart E. Multispot, multianalyte, immunoassay. Ann Biol Clin. 1990;48:655–66.

Ellington AA, Kullo IJ, Bailey KR, Klee GG. Antibody-based protein multiplex platforms: technical and operational challenges. Clin Chem. 2010;56:186–93. doi:10.1373/clinchem.2009.127514.

Gallagher RI, Silvestri A, Petricoin 3rd EF, et al. Reverse phase protein microarrays: fluorometric and colorimetric detection. Methods Mol Biol. 2011;723:275–301. doi:10.1007/978-1-61779-043-0_18.

Gao WM, Kuick R, Orchekowski RP, et al. Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. BMC Cancer. 2005;5:110. 1471-2407-5-110 [pii]. doi:10.1186/1471-2407-5-110.

Ghevaria N, Visser M, Hoffmann R. Quality control for a large-scale study using protein arrays and protein beads to measure immune response in serum and plasma. Proteomics. 2012;12:2802–7. doi:10.1002/pmic.201200082.

Gibson DS, Qiu J, Mendoza EA, et al. Circulating and synovial antibody profiling of juvenile arthritis patients by nucleic acid programmable protein arrays. Arthritis Res Ther. 2012;14:R77. doi:10.1186/ar3800.

Glokler J, Angenendt P. Protein and antibody microarray technology. J Chromatogr B Analyt Technol Biomed Life Sci. 2003;797:229–40. S1570023203006962 [pii].

Gonzalez-Gonzalez M, Jara-Acevedo R, Matarraz S, et al. Nanotechniques in proteomics: protein microarrays and novel detection platforms. Eur J Pharm Sci. 2012;45:499–506. doi:10.1016/j.ejps.2011.07.009.

Gutmann O, Kuehlewein R, Reinbold S, Niekrawietz R, Steinert CP, de Heij B, Zengerle R, Daub M. Fast and reliable protein microarray production by a new drop-in-drop technique. Lab Chip. 2005;5(6):675–81. Epub 2005 Apr 27.

Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 1999;19:1720–30.

Hanash S. Disease proteomics. Nature. 2003;422:226–32. doi:10.1038/nature01514.

Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. Nature. 2008;452:571–9. nature06916 [pii]. doi:10.1038/nature06916.

He M, Stoevesandt O, Palmer EA, et al. Printing protein arrays from DNA arrays. Nat Methods. 2008;5:175–7. doi:10.1038/nmeth.1178.

Huang RP, Huang R, Fan Y, Lin Y. Simultaneous detection of multiple cytokines from conditioned media and patient's sera by an antibody-based protein array system. Anal Biochem. 2001;294:55–62. S0003-2697(01)95156-5 [pii]. doi:10.1006/abio.2001.5156

Hunt I. From gene to protein: a review of new and enabling technologies for multi-parallel protein expression. Protein Expr Purif. 2005;40(1):1–22.

Kellar KL, Kalwar RR, Dubois KA, et al. Multiplexed fluorescent bead-based immunoassays for quantitation of human cytokines in serum and culture supernatants. Cytometry. 2001;45:27–36. doi:10.1002/1097-0320(20010901)45:1<27::AID-CYTO1141>3.0.CO;2-I [pii].

Kellar KL, Mahmutovic AJ, Bandyopadhyay K. Multiplexed microsphere-based flow cytometric immunoassays. Curr Protoc Cytom. 2006;Chapter 13:Unit13 1. doi:10.1002/0471142956.cy1301s35.

Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S. Cell-free production and stable-isotope label-ing of milligram quantities of proteins. FEBS Lett. 1999;442(1):15–9.

Kusnezow W, Jacob A, Walijew A, Diehl F, Hoheisel JD. Antibody microarrays: an evaluation of production parameters. Proteomics. 2003;3(3):254–64.

LaBaer J, Ramachandran N. Protein microarrays as tools for functional proteomics. Curr Opin Chem Biol. 2005;9:14–9.

Larsson K, Wester K, Nilsson P, et al. Multiplexed PrEST immunization for high-throughput affinity proteomics. J Immunol Methods. 2006;315:110–20. doi:10.1016/j.jim.2006.07.014.

Manzano-Roman R, Diaz-Martin V, Gonzalez-Gonzalez M, et al. Self-assembled protein arrays from an *Ornithodoros moubata* salivary gland expression library. J Proteome Res. 2012;11:5972–82. doi:10.1021/pr300696h.

Matarraz S, Gonzalez-Gonzalez M, Jara M, et al. New technologies in cancer. Protein microarrays for biomarker discovery. Clin Transl Oncol. 2011;13:156–61.

McQuain MK, Seale K, Peek J, Levy S, Haselton FR, McQuain MK, Seale K, Peek J, Levy S. Effects of relative humidity and buffer additives on the contact printing of microarrays by quill pins. Anal Biochem. 2003;320(2):281–91.

Merbl Y, Kirschner MW. Protein microarrays for genome-wide posttranslational modification analysis. Wiley Interdiscip Rev Biol Med. 2011;3:347–56. doi:10.1002/wsbm.120.

Miersch S, et al. Serological autoantibody profiling of type 1 diabetes by protein arrays. J Proteomics. 2013;94:486–96.

Mijakovic I, Macek B. Impact of phosphoproteomics on studies of bacterial physiology. FEMS Microbiol Rev. 2012;36(4):877–92.

Molero C, Rodríguez-Escudero I, Alemán A, et al. Addressing the effects of Salmonella internalization in host cell signaling on a reverse-phase protein array. Proteomics. 2009;9:3652–65.

Montor WR, Huang J, Hu Y, et al. Genome-wide study of Pseudomonas aeruginosa outer membrane protein immunogenicity using self-assembling protein microarrays. Infect Immun. 2009;77:4877–86. doi:10.1128/IAI.00698-09.

Paweletz CP, Charboneau L, Bichsel VE, et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene. 2001;20:1981–9. doi:10.1038/sj.onc.1204265.

Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12(1):32–42.

Poetz O, Ostendorp R, Brocks B, et al. Protein microarrays for antibody profiling: specificity and affinity determination on a chip. Proteomics. 2005;5:2402–11. doi:10.1002/pmic.200401299.

Ramachandran N, Raphael JV, Hainsworth E, et al. Next-generation high-density self-assembling functional protein arrays. Nat Methods. 2008a;5:535–8. doi:10.1038/nmeth.1210.

Ramachandran N, Srivastava S, Labaer J. Applications of protein microarrays for biomarker discovery. Proteomics Clin Appl. 2008b;2:1444–59. doi:10.1002/prca.200800032.

Rusmini F, Zhong Z, Feijen J. Protein immobilization strategies for protein biochips. Biomacromolecules. 2007;8:1775–89. doi:10.1021/bm061197b.

Schwenk JM, Lindberg J, Sundberg M, et al. Determination of binding specificities in highly multiplexed bead-based assays for antibody proteomics. Mol Cell Proteomics. 2007;6:125–32. T600035-MCP200 [pii]. doi:10.1074/mcp.T600035-MCP200.

Schwenk JM, Gry M, Rimini R, et al. Antibody suspension bead arrays within serum proteomics. J Proteome Res. 2008;7:3168–79. doi:10.1021/pr700890b.

Spurrier B, Ramalingam S, Nishizuka S. Reverse-phase protein lysate microarrays for cell signaling analysis. Nat Protoc. 2008;3:1796–808. nprot.2008.179 [pii]. doi:10.1038/nprot.2008.179.

Sreekumar A, Nyati MK, Varambally S, et al. Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. Cancer Res. 2001;61:7585–93.

Stoevesandt O, Taussig MJ. Affinity reagent resources for human proteome detection: initiatives and perspectives. Proteomics. 2007;7:2738–50. doi:10.1002/pmic.200700155.

Tannapfel A, Anhalt K, Hausermann P, et al. Identification of novel proteins associated with hepatocellular carcinomas using protein microarrays. J Pathol. 2003;201:238–49. doi:10.1002/path.1420.

Tao SC, Zhu H. Protein chip fabrication by capture of nascent polypeptides. Nat Biotechnol. 2006;24:1253–4. nbt1249 [pii]. doi:10.1038/nbt1249.

Templin MF, Stoll D, Schrenk M, et al. Protein microarray technology. Drug Discov Today. 2002;7:815–22.

Thanawastien A, Montor WR, Labaer J, et al. Vibrio cholerae proteome-wide screen for immunostimulatory proteins identifies phosphatidylserine decarboxylase as a novel Toll-like receptor 4 agonist. PLoS Pathog. 2009;5:e1000556. doi:10.1371/journal.ppat.1000556.

Theilacker N, Roller EE, Barbee KD, et al. Multiplexed protein analysis using encoded antibody-conjugated microbeads. J R Soc Interface. 2011;8:1104–13. rsif.2010.0594 [pii]. doi:10.1098/rsif.2010.0594.

Wong J, Sibani S, Lokko NN, et al. Rapid detection of antibodies in sera using multiplexed self-assembling bead arrays. J Immunol Methods. 2009;350:171–82. doi:10.1016/j.jim.2009.08.013.

Yu X, Schneiderhan-Marra N, Joos TO. Protein microarrays and personalized medicine. Ann Biol Clin (Paris). 2011;69:17–29. doi:10.1684/abc.2010.0512.

Chapter 9
Clinical Bioinformatics: A New Emerging Science of Biomarker Development

Xiaodan Wu, Xiaocong Fang, Zhitu Zhu, and Xiangdong Wang

9.1 Introduction

Cancer has become the leading cause of death in the last 50 years. Patients with early detection of cancer have better rate of the recovery and survival than patients with more advanced cancer. More than 90 % 5-year survival rate was found to associated with the detection of cancers at the stage one (Etzioni et al. 2003), which need cancer-specific and sensitive biomarkers to diagnose and monitor timely therapeutic interventions (Ullah and Aatif 2009). US Food and Drug Administration has approved a few biomarkers for early detection or screening of cancers, like prostate-specific antigen for prostate cancer, nuclear matrix protein 22 for bladder cancer, etc. Biomarkers can play roles before cancer diagnosis in risk assessment and screening, at diagnosis in classification, stage and grade, and after diagnosis in predicting response to therapy and toxicity related to treatment, selecting additional therapy and detecting recurrence (Fig. 9.1) (Ludwig and Weinstein 2005). Predictive biomarkers allow clinicians to assess clinical effects of chemotherapy and molecular targeted agents on response rate and survival time (Saijo 2012). The patients with poor responses have severe toxicities of chemotherapy or high prices of targeted drugs, if they did not have a reliable biomarker. Massive efforts have been carried out to identify such predictive biomarkers, of which some have been used in clinical trials. For example, somatic mutations in the tyrosine kinase domain of the epidermal growth factor receptor (EGFR) were shown to be a predictive marker for better efficacy of Gefitinib in patients with non small cell lung cancer (Lynch et al. 2004; Paez et al. 2004). Decisions in breast cancer are based on tumor size, node status, histological grade, age, estrogen receptor status, and EGF receptor 2 (HER2) status, of which two receptors have been considered as more important markers for the malignancy (Roukos 2010).

X. Wu • X. Fang • X. Wang (✉)
Department of Pulmonary Medicine, Zhongshan Hospital, Shanghai Respiratory Research Medicine, Fudan University, Shanghai, China
e-mail: physicianwuxd@126.com; fangxiaocong@gmail.com; xiangdong.wang@clintransmed.org

Z. Zhu (✉)
Department of Oncology, Liaoning Medical University Hospital, Jinzhou, China
e-mail: zhuzhitu@163.com

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*, Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_9

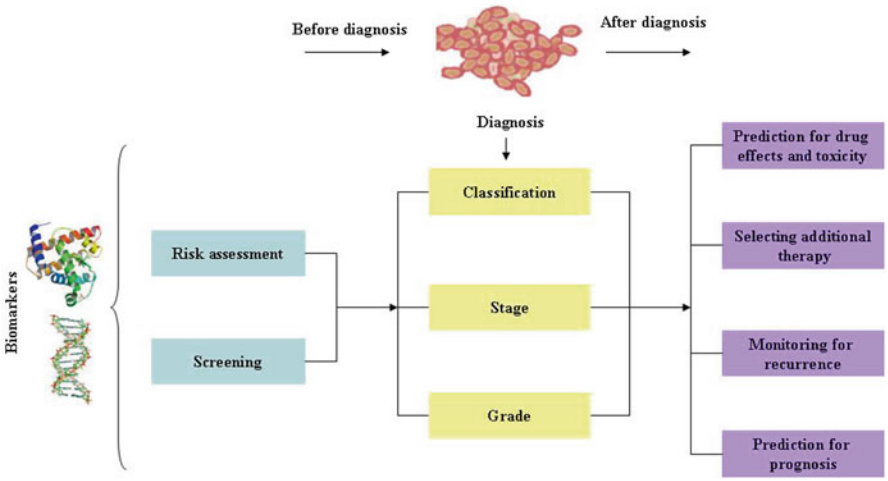


Fig. 9.1 Biomarkers can play roles before cancer diagnosis, at diagnosis and after diagnosis

However, cancer is a progressive and complex disease that results from the accumulation of multiple mutations, and an individual biomarker usually provides limited insights into cellular mechanisms that underlie tumorigenesis (Nibbe et al. 2010). It is important to explore the interactions between cancer-associated and/or specific genes and proteins, to understand how those genes and proteins work together as whole biological and clinical systems. The better understanding of gene, protein, and cell interactions and connections during the disease progression can benefit the identification of disease-specific biomarkers, pathways and targets for drug development (Barabasi et al. 2011). Recent studies focus on identification of network biomarkers, i.e. functionally associated biomarkers, which may coordinate changes during cancer development, progression or treatment (Deng et al. 2007; Jin et al. 2009; Vilar et al. 2009; Wang et al. 2009; Wang and Chen 2011). The most challenge is to translate such network biomarkers from the discovery of disease-specific biomarkers monitoring the functional integrity of the network perturbed by the diseases and defining better disease classification paving the way to personalized therapies (Barabasi et al. 2011). The present chapter aims to introduce new concepts for understanding and developing network biomarkers like clinical bioinformatics and systems clinical medicine, define the differentiation among biomarkers, network biomarkers, and dynamic network biomarkers, summarize applied methodologies in the discovery and development of biomarkers and biomarker networks, and present the new system to translate clinical descriptive information into clinical informatics in the digital form in order to integrate clinical phenotypes with bioinformatics findings.

9.2 Significance of Clinical Bioinformatics

Bioinformatics is mainly used to entail the creation and advancement of databases, algorithms, computational and statistical techniques, and theories and to solve problems generated from the analysis of biological data. However, the traditional bioinformatics have focused on the biomedical informatics rather than combining with the clinical information. There is a great and hard barrier to understand the association and specificity between bioinformatics and clinical information. There are increasing recognitions and growing needs to bridge the connection of bioinformatics analyses with clinical phenotypes.

Clinical bioinformatics is such a new emerging science to focus on the combination of clinical symptoms and signs with human tissue-generated bioinformatics and to get deep and full understanding of the risk factors, pathogenesis and progress of human diseases (Wang and Liotta 2011). The term "Clinical bioinformatics" was defined as "clinical application of bioinformatics-associated sciences and technologies to understand molecular mechanisms and potential therapies for human diseases", a new and important concept for the development of disease-specific biomarkers, mechanism-oriented understanding, and individualized medicine.

9.3 Introduction of Systems Clinical Medicine

Systems biology is used to investigate what the interactions between the components of biological systems are and how those interactions involve the function and behavior of the system (Laubenbacher et al. 2009). Systems biology is to integrate multi-factors, elements, methodologies, sources, and/or information together (Denis 2006). Dr. von Bertalanffy, one of the precursors of systems biology, mentioned that all existed systems should share the common property composed of interlinked components, where the similarities are shared in detailed structures and control designs (Trewavas 2006). Drs Hodgkin and Huxley, British neurophysiologists and Nobel prize winners, constructed a mathematical model to explain the developing potential of the neuronal cell propagation along the axon in 1955 (Hodgkin and Huxley 1952). The initial study of systems biology entitled "Systems Theory and Biology" was reported in 1966 (Mesarovic 1968). Systems biology was then used in various genome projects with the large increase in data from the omics, high-throughput experiments, and bioinformatics.

The treatment of complex diseases needs information on molecular mechanisms, pathogenesis, pathophysiology, developing processes, and potentials associated with to clinical outcomes. For this reason, systems biology is widely and increasingly applied to provide a systems-level understanding of the complex interactions between genes, proteins, and metabolites in the disease (Kitano 2002; Westerhoff and Palsson 2004). Recent studies in cellular immunology, molecular biology, genomics, transcriptomics, proteomics, or others, generated a large number of possible biomarkers to diagnose the disease and predict therapeutic response or prognosis of the diseases (Tumani et al. 2009). The systems biology is used to discover and develop biomarkers through the integration of the molecular data generated by omics studies with disease pathogenesis, signaling pathways, and biological networks. It is important to translate such information on different levels of biological complexity (genes, molecules, cells, tissues, and the organisms) to the physiological explanation of the clinical phenotypes and findings (Baranzini 2006). However, most of those biomarkers have the limit and barrier to be used in the clinical practice due to the lack of the proper validation, disease specificity, and association with clinical phenotypes. Those biomarkers may have the significance from the statistical analysis, while lack available and repeatable data to the dependence of clinical variants (Ioannidis et al. 2009; Ioannidis and Panagiotou 2011). Therefore, the process of biomarker discovery to find molecules associated with a given disease phenotype is the initial step, while there are a large number of validation to define the association of biological pathways with disease pathogenesis, course, and interventions. Thus, the systems clinical medicine should be considered and defined as an extension of systems biology and systems biomedicine to integrate the discovery, identification, validation, and development of gene-, protein-, cell-, and organ-based data and bioinformatics, with disease history, phenotypes (e.g. symptoms, signs, imaging, biochemical changes, functional tests, and therapies), responses, outcomes, and prognosis. The systems clinical medicine should be a new emerging science for understanding the disease, identifying disease-specific biomarkers, or developing new therapies.

9.4 Biomarkers, Network Biomarkers, and Dynamic Network Biomarkers

The biomarker has been considered as a measurable character to indicate biologic, pathogenic, pathophysiological, or pharmacologic processes and/or responses to the therapy (Biomarkers Definitions Working Group 2001; Simon 2005). There is a rapid progress for the combined application of omic science and computational technologies in the detection of genetic, proteomic and metabolomic biomarkers. The great increase in omics-related research produced a tremendous amount of information related to molecular markers (Hogeweg and Hesper 1978; Lau and Chiu 2009; Spencer et al. 2009). Various powerful data mining and statistical bioinformatics methods have been propagated to identify, prioritize, and classify robust and general biomarkers with high discriminatory ability (Baumgartner et al. 2011). Some online bioinformatics data libraries, such as Enzyme, KEGG, Gene Ontology, NCBI Taxonomy, SwissProt and TrEMBL were generated for the store and management of data.

Network biomarkers or dynamic network biomarkers were investigated and developed as new type of biomarkers that emphasize the relationship between

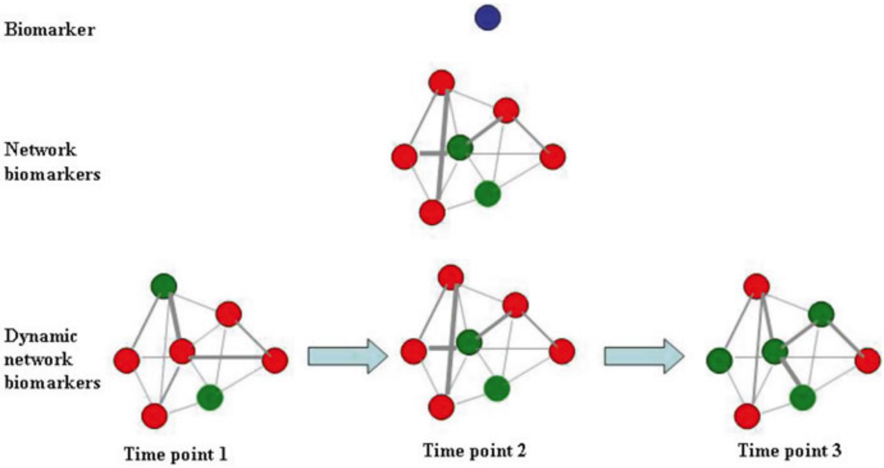


Fig. 9.2 The difference among biomarker, network biomarkers and dynamic network biomarkers

genes, proteins or metabolites (Jin et al. 2008; Wang 2011). Disease-specific biomarker/molecule networks can be build on basis on the knowledge of the genes, proteins, metabolites and their unparalleled expression levels in different conditions (Barabasi and Oltvai 2004). Cytoscape is one of the powerful softwares to be used to complete the network construction. iCTNet is a plugin for Cytoscape, built on a complex database that integrates interactions among human phenotypes, proteins, tissues and drugs. The relationships between diseases and genes, DNA, or proteins, and/or between therapeutic drugs and their targets were joined into a common database and visualized together in a network environment (Wang et al. 2011).

Dynamic network biomarkers show alterations of network biomarkers which are monitored and evaluated at different stages and severities during the development of diseases (Wang 2011). Biomarker provides one-dimensional information, while network biomarkers provide two-dimensional information by adding interactions of biomarkers. Dynamic network biomarkers provide a three-dimensional image of biomarker-biomarker interactions, not only by demonstrating the location and time of altered biomarkers, but also by showing time-dependent stronger or weaker interactions among biomarkers in the network (Wang 2011). The differences among biomarker, network biomarker and dynamic network biomarkers are shown in Fig. 9.2.

9.5 Applied Methodologies

A number of computational programs have been developed to generate and study disease-related networks and network biomarkers. Those methodologies include gene regulatory network inference tool (GRNInfer), gene regulatory network

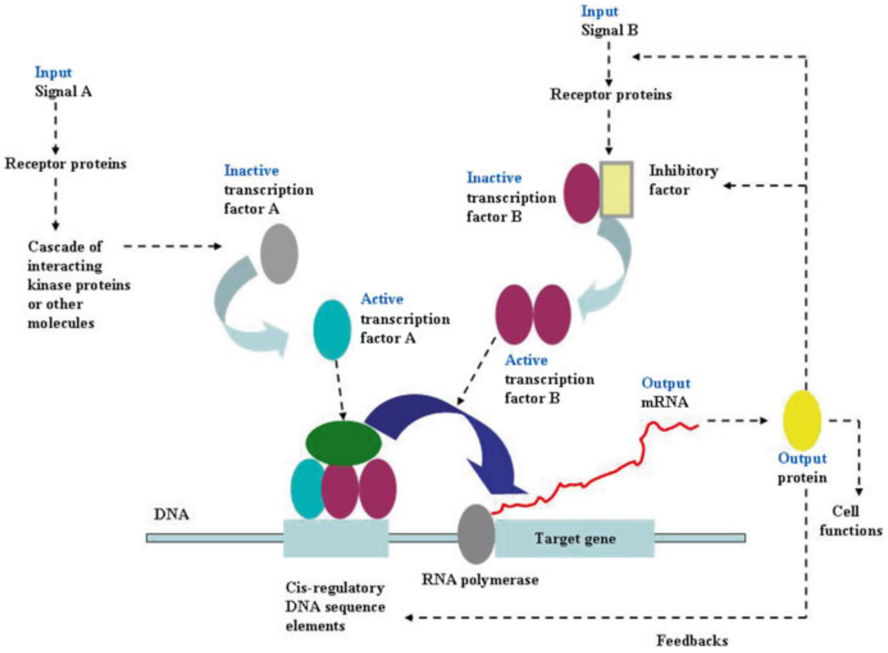


Fig. 9.3 Gene regulatory network

reconstruction tool with compound targets (nGNTInfer), inferring transcriptional regulatory networks from high-throughput data (nTRNInfer), inferring protein-protein interactions by parsimony principle (nInferPPI), inferring protein-protein interactions based on multi-domain cooperation (nMDCinfer), molecular network aligner (nMNAligner), detecting drug targets in metabolic networks by integer linear programming (nMetaILP), protein structure alignment tool based on multiple objective optimization (nSamo), annotating genes with positive samples (nAGPS), parsimonious tree grow method for haplotype inference (nPTG), identifying differentially expressed pathways via a mixed integer linear programming model (nMILPs), protein- RNA binding site prediction (nPRNA), or network ontology analysis (nNOA) (Wang 2011) (Fig. 9.3).

Methodologies for computational analysis can vary widely according to the question posed and the experimental data at hand, from highly abstracted models with correlative regression to highly specified models with differential equations, network component interaction, and logic modeling techniques (Kreeger and Lauffenburger 2010). A number of reviews have provided the discussion of those methods appropriate for various kinds of studies, outlining their respective strengths and weaknesses with respect to different applications (Aldridge et al. 2006; Cho et al. 2006; Papin et al. 2005).

9.6 Development and Application of DESS

Digital Evaluation Score System (DESS) is a new system for translating medical information into clinical informatics in patients (Chen et al. 2012a, b). With development of bioinformatics, there is a growing need to associate the clinical information with those high-throughput data. DESS was designed as a score system in order to combine clinical and biological information and validate the system in various stages and severities of the disease. DESS evaluate patient history, symptoms, signs and laboratory findings in the disease. DESS is less affected by individuals and may provide a new concept to interpret clinical data into computational language.

For example, acute respiratory distress syndrome (ARDS) affects a large number of patients with a poor prognosis, as an acute life-threatening form of hypoxemic respiratory failure characterized by increased pulmonary capillary permeability and edema. ARDS frequently complicates the clinical course of dysfunction of other organs or tissue including fluid and electrolyte imbalance, liver dysfunction or heart failure. However, there are no systemic score system for predicting the severity and outcomes of the disease.

A large amount of data were acquired and growing knowledge of diseases were obtained due to the application of high-throughput methodologies. Increased levels of chemokines, receptors, or proteins in plasma or bronchoalveolar lavage fluid were identified as biologic markers for the prediction of clinical outcomes, such as the receptor for advanced glycation end products (Calfee et al. 2008), plasma surfactant protein levels (Eisner et al. 2003), TNF receptor (Parsons et al. 2005), IL-8 and SP-D (Ware et al. 2011), alveolar granulocyte colony-stimulating factor and alpha-chemokines (Wiedermann et al. 2004). The predictive value of these inflammatory mediators for the outcome of patients with ARDS has been reported, but the results are inconclusive (Agouridakis et al. 2002; Kiehl et al. 1998). In addition, clinical information in traditional descriptive way was not able to compile with those “omics” data. Previous assessments of monitoring and predictive power for clinical practice, are graded with partial variables which are far less enough for high-throughput data. Digitalizing and essential clinical profiles would provide better way to bridge the gap between clinical and biological information. There is an urgent need to translate clinical messages into digitalized information. The emerging of clinical bioinformatics offers a new opportunity for understanding this problem by associating multidisciplinary knowledge.

The combination of biologic and clinical risk factors would provide a superior prognostic index for mortality in patients with ALI/ARDS, as compared with either biologic or clinical risk factors alone. We developed a multidimensional grading system to assess symptoms, signs, and laboratory findings, categorize the illness, and provide an easy way to relate clinical and biological data. A score index which included symptoms, signs and partial laboratory results was designed to assess patients with ARDS. Most common symptoms and signs in pulmonary diseases were involved as variables. Laboratory tests which reflect the function of vital organs and tissue were included. The various components of the index were assigned as different weights, as 0, 1, 2 and 4 (Tables 9.1, 9.2, and 9.3). Total points for each

Table 9.1 Variables and point values used for new score system (history and risk factors)

Variables	Points			
	0	1	2	4
General condition				
Age (years)	<30	30–50	51–70	>70
Posture for admission	Walk	Walk with support	On a wheelchair	Recline on a bed
History				
Cough	No	Yes		
Cough severeness	No	≤1 week	1–2 weeks	≥2weeks
Sputum	No	White, and small amount	White, relatively larger amount	Yellow
Short breathness	No	Only under severe activity	In daily activity	At rest
Hemoptysis	No	A little	Median	Large
Fever (°C)	No	37.3–38	38.1–39	≥39
Duration of fever	No	≤1 week	1–21 weeks	≥21 weeks
Chest pain	No	Under severe activity	Under daily activity	At rest
Headache or dizzy	No	Yes		
Feel weak	No	Yes		
Limitation of activity	No	Mild	Marked	Severe
Orthopnea	No			Yes
Chill	No			Yes
Cold sweat	No			Yes
Abdominal pain	No			Yes
Nausea and vomiting	No			Yes
Appetite	Good	Semi-liquid diet	Liquid diet	Absolute diet
Stool	Normal			Abnormal
Urine	Normal	Decreased urine volume but ≥ 500 mL/24 h	Oliguria	Anuira
Risk factors				
Smoking (park year)	5	6–10	11–20	>20
Asthma	No	≤10 years	10–20 years	≥20 years
Hypertension	No	≤5 years	5–10 years	≥10 years
Diabetes mellitus	No	≤5 years	5–10 years	≥10 years
Hyperlipoidemia	No	≤5 years	5–10 years	≥10 years
Emphysema	No	≤10 years	10–20 years	≥20 years
Chronic bronchitis	No	≤10 years	10–20 years	≥20 years
Chronic obstructive pulmonary disease (COPD)	No	≤10 years	10–20 years	≥20 years
Coronary heart disease	No			Yes

(continued)

Table 9.1 (continued)

Variables	Points			
	0	1	2	4
Arrhythmia	No			Yes
Operation history	No	1 year before	3 months to 1 year	Within 3 months
Gastrointestinal bleeding	No	1 year before	3 months to 1 year	Within 3 months
Pancreatitis	No	1 year before	3 months to 1 year	Within 3 months
Cholecystitis	No	1 year before	3 months to 1 year	Within 3 months
Pneumonia	No	1 year before	3 months to 1 year	Within 3 months
Appendicitis	No	1 year before	3 months to 1 year	Within 3 months
Tumor	No	1 year before	3 months to 1 year	Within 3 months

Table 9.2 Variables and point values used for new score system (signs)

	Points			
Variables	0	1	2	4
Signs				
Nutrition	Good	Median		Poor or overweight
Consciousness	<Conscious	Hypersomnia	Confusion	Coma
Temperature (°C)	<37.3	37.3–38	38.1–39	≥39.1
Heart rate (beat/min)	60–100			>100, or<60, or with any kind of arrhythmia
Respiratory (rate/min)	16–18	19–20, or 12–15	21–24, or 8–11	>24, or<8
Blood pressure (mmHg)	Diastolic: 120–140	Diastolic: 140–159 or 100–119	Diastolic: 160–179 or 80–99	Diastolic: ≥180 or ≤80
	Systolic 80–90	Systolic 90–99 or 60–79	Systolic 100–109 or 40–59	Systolic ≥110 or≤40
Cyanosis	No			Yes
Complexion	Normal		Flush	Pale
Jaundice	No			Yes
Enlargement of lymph nodes	No			Yes
Three depression sign	No			Yes
Barrel chest	No			Yes
Chest palpitation	Negative			Positive signs
Chest percussion	Negative			Positive signs
Rales	No	Single side<1/3 area	Single side1/3–1/2, or bilateral <1/3	Single side>1/2, or bilateral>1/3
Heart examination	Negative			Positive signs
Peripheral vascular sign	Negative			Positive signs
Edema of lower limbs	No			Yes
Abdominal examination	Negative			Positive signs
Pathologic signs for nervous system	Negative			Positive signs

Table 9.3 Variables and point values used for new score system (laboratory tests and imaging)

	Points			
Variables	0	1	2	4
Laboratory tests				
Hemoglobin (g/L)	Male:120–160	90-lower limit of normal	60–90	<60
	Female:110–150			
WBC ($\times 10^9$/L)	4–10	1.5–4	<1.5	>10, or <1.5
Neutrophil percentage (%)	50–70			>70, or <50
Lymphocyte percentage (%)	20–40			<20, or >40
Platelet ($\times 10^9$/L)	100–300		>300, or < 100	
Urine protein	Negative	+	++	+++
Urine glucose	Negative	+	++	+++
Urine ketones	Negative	+	++	+++
Urine WBC	Negative	+	++	+++
Urine RBC	Negative	+	++	+++
Occult blood	Negative	+	++	+++
Fecal WBC	Negative	+	++	+++
Total protein (g/L)	66–87	50–66	40–50	<40
Albumin (g/L)	35–55		28–35	<28
Bilirubin (μmol/L)	<34.2	34.2–171	171–342	>342
ALT (U/L)	<30			>30
AST (U/L)	<50			>50
ALP (U/L)	Within normal range			Beyond
Gamma-GT	Within normal range			Beyond
Urea (mmol/L)	2.5–7.1	7.1–9	40,441	>20
Creatinine (μmol/L)	40–120	120–150	150–200	>200
Cholesterol (mmol/L)	3.1–5.9	5.9–7	40,367	>8
GFR (mL/(min × 1.73 m^2))	≥90	60–89	30–59	15–29
Cholesterol	3.1–5.9	6.0–7.0	7.1–8.0	>8.0
Triglyceride (mmol/L)	0.6–2.0	2.1–3.0	3.1–4.0	>4.0
HDL (mmol/L)	1.03–2.07	0.91–1.03		≤0.91
LDL (mmol/L)	≤3.12	3.12–3.16	3.16–3.64	>3.64
cTnT (ng/mL)	<0.03	0.03–0.3	0.3–3	>3
CK (U/L)	26–140			>140
Ck-MB (U/L)	0–23			>23
NT-proBNP (pg/mL)	0–300	301–2,000	2,001–5,000	>5,000
Left ventricular ejection fraction, LVEF (%)	>60	40–60	30–39	<30
Na (mmol/L)	135–145	146–155, or 125–134	156–165, or 115–124	>165, or <115
K (mmol/L)	3.5–5.5	3–3.4	2.5–2.9	>5.5, or < 2.5

(continued)

Table 9.3 (continued)

	Points			
Variables	0	1	2	4
Cl (mmol/L)	95–105			<95, or > 105
Ca (mmol/L)	2.25–2.58			>2.58, or < 2.25
P (mmol/L)	0.97–1.61			>1.61, or < 0.97
pH	7.35–7.45			>7.45, or < 7.35
PaO_2 (mmHg)	≥90	60–90	40–60	<40
$PaCO_2$ (mmHg)	35–45	45–50		>50
SaO_2 (%)	≥90	80–90	60–80	<60
D Dimer (mg/L)	0.02–0.8	0.9–2.0	2.1–5.0	>5.0
INR	0.5–1.2	1.21–2.0	2.1–3.0	>3.0
Prothrombin time prolonged (s)	0	Within 4 s	4–6	>6
Fasting blood glucose (mmol/L)	<5.8	5.8–7		>7
Glycosylated hemoglobin, HbA1c (%)	40,274	40,337	40,399	>9
Increased numbers of tumor marker	0	1–2	2–4	≥4
C-reactive protein, CRP (mg/L)	≤10	40,481	30–90	>90
Anti “O” antibody	–	+		
ENA	–	+		
ANA	–	+		
Lung imaging				
Lung consolidation	No	Single lobe < 1/3 area	Single lobe 1/3–1/2	Single lobe > 1/2, or bilateral
Enlargement of lymph nodes	No			Yes
Pleural effusion	No	Single lobe < 1/3 area	Single lobe 1/3–1/2	Single lobe > 1/2, or bilateral
Emphysema	No			Yes
Pulmonary edema	No	Single lobe < 1/3 area	Single lobe 1/3–1/2	Single lobe > 1/2, or bilateral

index were added, so that our index ranged from 0 to 456 points, with higher scores indicating a severer condition. Comparisons between student group and physician group were performed by unpaired Student test (Table 9.4). The differences within groups were completed by one-way analysis of variance (Table 9.5).

DESS as a simple, easily calculable scoring model can be used monitor the severity of the disease. During the past decades, scoring systems based on physiologic abnormalities have been successful in measuring severity of illness among

Table 9.4 Comparison on single history between student group and physician group

	Points (mean ± SD)		
Patient	Students (n = 5)	Physicians (n = 5)	p-value
1	78.2 ± 9.33	85.6 ± 10.4	0.454
2	103.5 ± 5.65	110.7 ± 6.91	0.722
3	63.2 ± 6.01	62.8 ± 8.23	0.801
4	55.4 ± 10.32	51.2 ± 8.09	0.65
5	85.4 ± 12.4	92.2 ± 9.45	0.235

Table 9.5 Comparison performance within student and physician groups

Individual	Points (mean ± SD)	p-value
Student 1	67.2 ± 24.3	All >0.05
Student 2	73.3 ± 27.1	
Student 3	66.9 ± 22.19	
Student 4	75.3 ± 29.36	
Student 5	77.6 ± 20.1	
Physician 1	78 ± 24.05	
Physician 2	75.2 ± 26.56	
Physician 3	79.4 ± 16.35	
Physician 4	80.4 ± 23.43	
Physician 5	83.2 ± 22.7	

critically ill patients. Examples include Acute Physiology and Chronic Health Evaluation III and IV (Zimmerman et al. 2006), Simplified Acute Physiology Score II (Le Gall et al. 1993), and the Mortality Probability Model (Lemeshow et al. 1993). However, those existed systems were unstable in practical application and less accuracy of the developed clinical measurements. A simple clinical predictive index for objective estimates of mortality in acute lung injury was tested in 2009 (Cooke et al. 2009). This simple point score, incorporating age, 24-h fluid balance, hematocrit, and bilirubin, is able to discriminate patients with high mortality from those with a lower mortality. However, this model can be used to inform prognosis (e.g., in counseling patients or families) but should not be used for decision making.

Combining clinical and biologic markers for diagnosis and prediction of clinical outcomes has been of major value in many clinical disorders, including solid tumors, hematopoietic malignancies, and rheumatologic conditions. The largest studies integrating clinical and biologic predictors have been done in patients at risk for and with existing cardiovascular disease (Danesh et al. 2004). However, such prognostic model with a combination of both biologic markers and clinical risk factors for ARDS has not been established yet. More important, our index which combines variables by means of a simple scale may be applied with genomics and proteomics results. Previous assessments, such as BODE index (Celli et al. 2004), are of excellent monitor and predictive power for clinical practice. However they are graded with only partial variables which might not enough for high-throughput data.

Therefore, digitalizing essential clinical profiles, such as symptoms and signs, by questionnaires and/or scores, would provide the direct vision for physicians and shorten the distance between bioinformatics and clinical phenotypes. However, the DESS should be further validated, developed, and confirmed by other groups and centers, standardized and optimized with the programmed software, and has the disease specificity.

In conclusion, clinical and medical informatics as the part of clinical bioinformatics should be fully considered before and after any investigation of omics-based studies. It is important to have a special attention from omics scientists to explore the combination between advanced omics-based biotechnologies, clinical phenotypes, tissue imaging and profiling, and organ dysfunction score systems, to improve the clinical outcomes of these patients (Wang et al. 2006). The use of high-throughput techniques and computerized databases for gene and protein expression profiling has become a mainstay of biomedical research. Clinical bioinformatics could be achieved from the combination of clinical informatics, medical informatics, bioinformatics and informatics by collaborations among clinicians, bioinformaticians, computer scientists, biologists, and mathematicians (Chen et al. 2012a, b).

References

Agouridakis P, Kyriakou D, Alexandrakis MG, Prekates A, Perisinakis K, Karkavitsas N, Bouros D. The predictive role of serum and bronchoalveolar lavage cytokines and adhesion molecules for acute respiratory distress syndrome development and outcome. Respir Res. 2002;3:25.

Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK. Physicochemical modelling of cell signalling pathways. Nat Cell Biol. 2006;8:1195–203.

American Cancer Society. Cancer statistics, 2005. www.cancer.org

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.

Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.

Baranzini SE. Systems-based medicine approaches to understand and treat complex diseases. The example of multiple sclerosis. Autoimmunity. 2006;39:651–62.

Baumgartner C, Osl M, Netzer M, Baumgartner D. Bioinformatic-driven search for metabolic biomarkers in disease. J Clin Bioinform. 2011;1:2.

Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69:89–95.

Calfee CS, Ware LB, Eisner MD, Parsons PE, Thompson BT, Wickersham N, Matthay MA. Plasma receptor for advanced glycation end products and clinical outcomes in acute lung injury. Thorax. 2008;63:1083–9.

Celli BR, Cote CG, Marin JM, Casanova C, Montes DOM, Mendez RA, Pinto PV, Cabral HJ. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. N Engl J Med. 2004;350:1005–12.

Chen H, Song Z, Qian M, Bai C, Wang X. Selection of disease-specific biomarkers by integrating inflammatory mediators with clinical informatics in AECOPD patients: a preliminary study. J Cell Mol Med. 2012a;16:1286–97.

Chen H, Wang Y, Bai C, Wang X. Alterations of plasma inflammatory biomarkers in the healthy and chronic obstructive pulmonary disease patients with or without acute exacerbation. J Proteomics. 2012b;75:2835–43.

Cho CR, Labow M, Reinhardt M, van Oostrum J, Peitsch MC. The application of systems biology to drug discovery. Curr Opin Chem Biol. 2006;10:294–302.

Cooke CR, Shah CV, Gallop R, Bellamy S, Ancukiewicz M, Eisner MD, Lanken PN, Localio AR, Christie JD. A simple clinical predictive index for objective estimates of mortality in acute lung injury. Crit Care Med. 2009;37:1913–20.

Danesh J, Wheeler JG, Hirschfield GM, Eda S, Eiriksdottir G, Rumley A, Lowe GD, Pepys MB, Gudnason V. C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. N Engl J Med. 2004;350:1387–97.

Deng X, Geng H, Ali HH. Cross-platform analysis of cancer biomarkers: a Bayesian network approach to incorporating mass spectrometry and microarray data. Cancer Inform. 2007;3:183–202.

Denis N. The music of life: biology beyond the genome. Oxford: Oxford University Press; 2006. p. 176.

Eisner MD, Parsons P, Matthay MA, Ware L, Greene K. Plasma surfactant protein levels and clinical outcomes in patients with acute lung injury. Thorax. 2003;58:983–8.

Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, Radich J, Anderson G, Hartwell L. The case for early detection. Nat Rev Cancer. 2003;3:243–52.

Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol. 1952;117:500–44.

Hogeweg P, Hesper B. Interactive instruction on population interactions. Comput Biol Med. 1978;8:319–27.

Ioannidis JP, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. JAMA. 2011;305:2200–10.

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009;41:149–55.

Jin G, Zhou X, Wang H, Zhao H, Cui K, Zhang XS, Chen L, Hazen SL, Li K, Wong ST. The knowledge-integrated network biomarkers discovery for major adverse cardiac events. J Proteome Res. 2008;7:4013–21.

Jin G, Zhou X, Cui K, Zhang XS, Chen L, Wong ST. Cross-platform method for identifying candidate network biomarkers for prostate cancer. IET Syst Biol. 2009;3:505–12.

Kiehl MG, Ostermann H, Thomas M, Muller C, Cassens U, Kienast J. Inflammatory mediators in bronchoalveolar lavage fluid and plasma in leukocytopenic patients with septic shock-induced acute respiratory distress syndrome. Crit Care Med. 1998;26:1194–9.

Kitano H. Systems biology: a brief overview. Science. 2002;295:1662–4.

Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. Carcinogenesis. 2010;31:2–8.

Lau AT, Chiu JF. Biomarkers of lung-related diseases: current knowledge by proteomic approaches. J Cell Physiol. 2009;221:535–43.

Laubenbacher R, Hower V, Jarrah A, Torti SV, Shulaev V, Mendes P, Torti FM, Akman S. A systems biology view of cancer. Biochim Biophys Acta. 2009;1796:129–39.

Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA. 1993;270:2957–63.

Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA. 1993;270:2478–86.

Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer. 2005;5:845–56.

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;350:2129–39.

Mesarovic M. Systems theory and biology. Berlin: Springer; 1968.

Nibbe RK, Koyuturk M, Chance MR. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol. 2010;6:e1000639.

Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304:1497–500.

Papin JA, Hunter T, Palsson BO, Subramaniam S. Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol. 2005;6:99–111.

Parsons PE, Matthay MA, Ware LB, Eisner MD. Elevated plasma levels of soluble TNF receptors are associated with morbidity and mortality in patients with acute lung injury. Am J Physiol Lung Cell Mol Physiol. 2005;288:L426–31.

Roukos DH. Novel clinico-genome network modeling for revolutionizing genotype-phenotype-based personalized cancer care. Expert Rev Mol Diagn. 2010;10:33–48.

Saijo N. Critical comments for roles of biomarkers in the diagnosis and treatment of cancer. Cancer Treat Rev. 2012;38:63–7.

Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. J Natl Cancer Inst. 2005;97:866–7.

Spencer SJ, Bonnin DA, Deasy JO, Bradley JD, El NI. Bioinformatics methods for learning radiation-induced lung inflammation from heterogeneous retrospective and prospective data. J Biomed Biotechnol. 2009;2009:892863.

Trewavas A. A brief history of systems biology. "Every object that biology studies is a system of systems." Francois Jacob (1974). Plant Cell. 2006;18:2420–30.

Tumani H, Hartung HP, Hemmer B, Teunissen C, Deisenhammer F, Giovannoni G, Zettl UK. Cerebrospinal fluid biomarkers in multiple sclerosis. Neurobiol Dis. 2009;35:117–27.

Ullah MF, Aatif M. The footprints of cancer development: cancer biomarkers. Cancer Treat Rev. 2009;35:193–200.

Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. J Theor Biol. 2009;261:449–58.

Wang X. Role of clinical bioinformatics in the development of network-based biomarkers. J Clin Bioinform. 2011;1:28.

Wang YC, Chen BS. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. BMC Med Genomics. 2011;4:2.

Wang X, Liotta L. Clinical bioinformatics: a new emerging science. J Clin Bioinform. 2011;1:1.

Wang X, Adler KB, Chaudry IH, Ward PA. Better understanding of organ dysfunction requires proteomic involvement. J Proteome Res. 2006;5:1060–2.

Wang HQ, Wong HS, Zhu H, Yip TT. A neural network-based biomarker association information extraction approach for cancer classification. J Biomed Inform. 2009;42:654–66.

Wang L, Khankhanian P, Baranzini SE, Mousavi P. iCTNet: a Cytoscape plugin to produce and analyze integrative complex traits networks. BMC Bioinform. 2011;12:380.

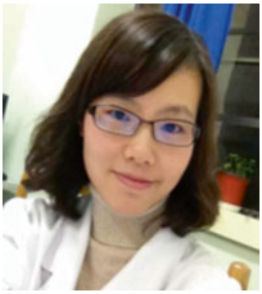
Ware LB, Koyama T, Billheimer DD, Wu W, Bernard GR, Thompson BT, Brower RG, Standiford TJ, Martin TR, Matthay MA. Prognostic and pathogenetic value of combining clinical and biochemical indices in patients with acute lung injury. Chest. 2011;137:288–96.

Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. Nat Biotechnol. 2004;22:1249–52.

Wiedermann FJ, Mayr AJ, Kaneider NC, Fuchs D, Mutz NJ, Schobersberger W. Alveolar granulocyte colony-stimulating factor and alpha-chemokines in relation to serum levels, pulmonary neutrophilia, and severity of lung injury in ARDS. Chest. 2004;125:212–19.

Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med. 2006;34:2517–29.

Xiaodan Wu, MD, graduated from Shanghai Medical College, Fudan University. Dr Wu works as in an attending physician at Department of Respiratory Medicine in Zhongshan Hospital. She majored in the research for COPD, sleep apnea and ARDS and won the golden award of European Respiratory Society in 2012.



Xiaocong Fang, MD, graduated from Shanghai Medical College, Fudan University. Dr Fang works as in an attending physician at Department of Respiratory Medicine in Zhongshan Hospital. She majored in the research for COPD.

Zhitu Zhu, MD, Ph.D., graduated from Dalian Medical University and gained a doctorate in clinical oncology in China Medical University. Dr Zhu works as the professor of oncology, a chief physician, and deputy president of Liaoning Medical University Hospital. He was honored as youth talent of Liaoning Medical University in 2011.

He also serves as the standing member of scientific organizations, including board directors of the Society of Medical Biology Immune, Chinese Medical Association; professional committee of chemotherapy, Chinese Pharmacological Society; the standing committee of tumor biological therapy professional committee in Liaoning province; the standing committee of a professional committee of chemotherapy and tumor metastasis, Liaoning anti-cancer association; the gastric cancer professional committee in Life Science Institute of Liaoning Province; and the lymphoma professional committee of Liaoning anti-cancer association. Dr Zhu is the member of editorial board of Molecular and Cellular Therapies, and serves as the vice-chairman of Liaoning Province Health Information Society and a managing director of professional committee of tumor cell biology affiliated to Liaoning Institute of Cell Biology. He published more than 30 papers in cancer research core journals.

Xiangdong Wang, MD, Ph.D., is a distinguished professor of respiratory medicine at Fudan University, deputy director of Shanghai Respiratory Research Institute, adjunct professor of Clinical Bioinformatics at Lund University, and visiting professor of King's College of London. He serves as a Director of Biomedical Research Center, Fudan University Zhongshan Hospital. His main research is focused on clinical bioinformatics, disease-specific biomarkers, cancer immunology, and molecular and cellular therapies.

His group integrates clinical informatics with omics science and bioinformatics to identify and validate disease-specific biomarkers and therapeutic targets in chronic lung diseases and lung cancer. His group initially developed the mirror-butterfly chemical structure of phosphoinositide 3-kinase inhibitor to prevent and treat chronic lung inflammation and injury, in combination of his pharmaceutical experience of drug discovery and development.

In addition, Dr Wang serves as the Executive Vice President of International Society for Translational Medicine, Chairman of Executive Committee of International Society for Translational Medicine, Deputy President of Chinese National Professional Society of Insurance and Health and a senior advisor of Chinese Medical Doctor Association, and Director of National Program of Doctor-Pharmaceutist communication. Dr Wang was appointed as the principal scientist, global disease advisor, Medical Monitor and Director, and Chairman of Director Board in a number of pharmaceutical companies, e.g. Astra Draco, AstraZeneca, PPT and CatheWill. He worked on pharmacology profiles of target identification and validation, drug screening and optimization, drug PK and PD profile, and translation between discovery and development in areas of respiratory diseases, inflammation and cancer. He acted as the adjunct professor of Molecular Bioscience at North Carolina State University, the member of American Thoracic Society International Health Committee, USA, and is the author of more than 200 scientific publications.

Chapter 10
Rapid Advances in the Field of Epigenetics

Takeshi Kawamura

Abstract Epigenetic regulations have been known as phenomena such as the X chromosome inactivation since the mid-twentieth century. Recently, DNA methylation and post-translational modification on histones, which consist chromatin, have been proved to mediate the epigenetic regulations. Then, DNA methyltransferases and histone deacetylases have been implicated in carcinogenesis to be attractive drug targets.

More recently, the filed of epigenome is rapidly advancing by findings such as methyltransferases and demethylases for histones and DNA demethylases and about 40 % of the genome was transcribed into none-coding RNA which participated in further transcriptional regulations. It also has been known that many life phenomena are regulated epigenetically with familiar factors including vitamins, environmental hormones and viral infection.

In this chapter, Sect. 10.1 presents the history of epigenome, the epigenetic regulators including DNA, histones and RNA, and some epigenetic phenomena such as reprograming and differentiation. Section 10.2 describes epigenome drug targets and current situation of the drug development. In addition, as the latest approaches, we illustrate genomic next generation sequencer and proteomic mass spectrometer.

Mass spectrometry is now approaching to the level to identify almost all of the translated proteins and have advantages in direct detection of post-translational modifications by their molecular shift. As an example of the epigenome analyses, we describe the methods, challenges and perspectives to reveal combinatorial histone modifications.

Keywords Epigenetics • DNA • Sample preparation • Mass spectrometry • Histone • Quantitation

T. Kawamura, Ph.D. (✉)
Laboratory for Systems Biology and Medicine (LSBM), Research Center for Advanced Science and Technology (RCAST), The University of Tokyo, Tokyo, Japan
e-mail: kawamura@med.rcast.u-tokyo.ac.jp

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*, Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_10

Abbreviations

DNMT	DNA methyltransferase
HAT	Histone Acetyl Transferase
HDAC	Histone deacetylase
HMT	Histone methyltransferase
KDM	Lysine Demethylase
ncRNA	Non-coding RNA
SAM	S-adenosylmethionine
SAH	S-adenosylhomocysteine
HTS	High throughput screening
NGS	Next Generation DNA Sequencer
SNP	Single nucleotide polymorphism
LC-MS	Liquid Chromatography Mass Spectrometry
ETD	Electron transfer dissociation
ECD	Electron Capture Dissociation
SRM	Selected Reaction Monitoring
MRM	Multiple Reaction Monitoring

10.1 Epigenetic Regulation

10.1.1 What Is Epigenome?

Conrad Waddington coined the term epigenetics after epigenesis, a concept that establishing a phenotype must need some molecular events other than DNA information during development and early embryogenesis, in contrast to the classical gene-centric concept (Waddington 1942). The term very broadly refers to all molecular events that modulate gene expression to direct a desired phenotype. With molecular genetics growing rapidly, epigenetics is now more narrowly defined as a study of changes in gene expression that are heritable (Ng et al. 2010; Seong et al. 2011) without any changes in DNA sequence. Those changes can generally be mediated by structural alterations of chromosomal constituents such as covalent DNA modifications, histone variants and modifications, and non-coding RNA, which are collectively named “epigenome”, epi (above) + genome. Epigenomics aims to study how various DNA, RNA and protein components interact between one another and coordinately regulate gene functions, and one of the important fields in systems biology understanding a life as dynamic biological networks of those components. Epigenomics will rapidly be growing with the advent of new research fields and methodologies of the broad range, and will has a great potential to uncover an important biological network underlying a variety of biological processes with the aid of other omics knowledge (Fig. 10.1). Epigenetics leads to introducing a concept that acquired characters are inheritable, thereby impacting on the gene-centric Darwinian evolution theory. In the next section,

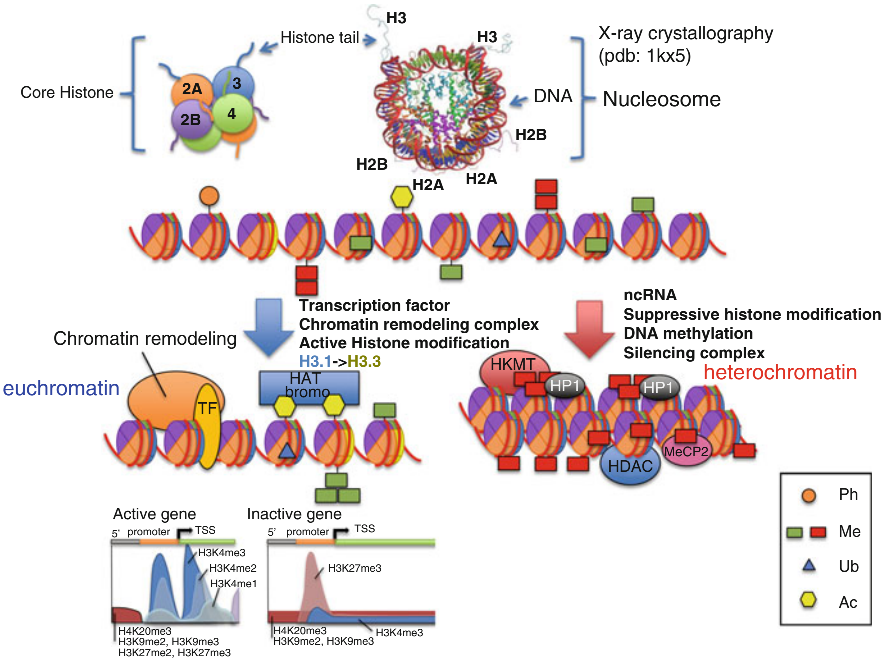


Fig. 10.1 An overview of epigenetic regulation (*Ph* phosphorylation, *Me* methylation, *Ac* acetylation, *Ub* ubiquitination) (David Allis et al. 2007; Kooistra and Helin 2012; Schones and Zhao 2008)

we are starting this chapter to explain typical and classical examples of epigenetic events, X-inactivation and genomic imprinting.

10.1.2 X-Inactivation and Genomic Imprinting

Before we know that the DNA is gene, the epigenetic phenomenon such as X-inactivation has been well known: For example, one of two X-chromosomes in mammals condensed into Barr Body, which can be identified under the light microscope, and then the transcription on the Barr body is under its repression (Kinoshita et al. 2008). This phenomenon influences not on DNA sequence but does on its gene expression on X chromosome. Similarly, genomic imprinting also takes place epigenetically. While an allele is randomly inactivated in X-inactivation, the genomic imprinting does inactivate a specific allele of those from father (paternal) or mother (maternal). A behavioral imprinting is derived from the class of long-term memory being evoked by short-term stimulus. This is well exemplified by the observation that a baby duck recognizes its mother by recognizing the first moving object in its front (Lorenz 1958). The genomic imprinting is associated with development, whereas the behavioral imprinting is equipped after birth.

Both the X-inactivation and genomic imprinting involve the inactivation in one of the paired paternal and maternal alleles where only one allele is transcribed into a protein. That is, one of the paired alleles is condensed into Barr body so as not to be transcribed but the other allele is transcribed. The X-inactivation occurs randomly within units of individual cells while the genomic imprinting takes place to units of individual genes. The X-inactivation occurs especially higher organisms. Different types of inactivation are seen in Eutheria and Marsupialia (with pouch): i.g. at random in Eutheria but on paternal allele in Marsupialia. Therefore, the X-inactivation in Marsupialia is much similar to genomic imprinting. In human, maternal genes are expressed in placenta due to X chromosome genomic imprinting during early developmental stage. It is most likely thought that the imprinting is deactivated and converted into X-inactivation during fetation (Reik 2007).

These phenomena seem to be driven from the gene-dosage compensation between male with only one maternal X-chromosome and female with two parental X-chromosomes. The genomic imprinting occurs not only on sex chromosomes but also on autosomal chromosomes in an invisible fashion under the light microscope unlike Barr body. For example, PEG10 gene on chromosome 7, which is essential in placenta formation, is expressed only from a paternal allele, and so placenta becomes to be hypoplastic when the paternal allele has a mutation gene and the maternal allele a normal gene. Numerous such genes have been reported so far (Dvash and Fan 2009; Lee and Bartolomei 2013).

As described above, although epigenetic regulation is defined as phenotypic change without DNA mutation, there are some easily confusable phenomena such as differences in immune system between twins. Even monozygotic twins have different antibodies. This is not because of epigenetic regulation but gene rearrangement, which is due to random rearrangement of genes that code antibodies during development.

On the other hand, one of well-known examples of epigenetic regulation in clone animals is the case of Calico cats (Shin et al. 2002). The fact that most of Calico cats are female exemplifies a sex-linked inheritance due to X-inactivation. The genes determining their fur color are coded on X-chromosome. The genes coding black and white fur are on autosomal chromosomes but that of brown is on X-chromosome. When brown gene is dominant (X^B), coat turns to brown, and when the gene is recessive (X^b), coat turns to black and white. Males have only one X-chromosome, so they turn into brown (X^BY) or black and white (X^bY). Females have two X-chromosomes and one of which is inactivated randomly cell by cell, and so they turn to brown ($X^BX^B = X^B +$ Barr body), black and white ($X^bX^b = X^b +$ Barr body), and tri-color ($X^BX^b = X^B +$ Barr body or $X^b +$ Barr body). Therefore, the clones of Calico cats ought to have the different patterns of black, brown and white furs.

Whereas such the X-inactivation is involved during development, similar epigenetic regulations are observed even after birth and universal in human life. These epigenetic regulations are stored from generation to generation, which can be referred not as genetic inheritance but as epigenetic inheritance. It had been previously believed that before birth epigenetic information is completely erased during gametogenesis when gametes fuse into a zygote. Such a mechanism is called “reprogramming”. It has been recently suggested that “reprogramming” is incomplete to allow the information so as to be partly inherited.

10.1.3 Structure of Chromatin and Epigenetic Modification

The essence for the epigenetic regulation of a gene expression is an involvement of transcription factors without DNA mutation. Transcription factors bind on the promoter region recruiting a RNA polymerase to activate transcription, which results in regulation of a gene expression. Such an activator sometimes can turn to be a repressor, which mostly depends on both a factor and its interacter to form a complex. It has been attempted to attain drug targets by analyzing such a complex. Transcription factors are previously considered to be activated by stimulus outside of the cells through cascades such as phosphorylation to bind to specific regions on DNA. However, now we know that the same stimulus can trigger a different gene activation including repression, depending on the types of cells, in which structural changes in the complexes of DNA and histones are largely involved.

It had been uncovered on the middle of twentieth century that a chromosome consists of both DNAs and histones, and that DNA is the gene and carries inheritable information. However, as the field of epigenetics has been widely expanded, a view about DNA has been currently changing, that is just a map describing all of the human body. Chromosomes are formed with the minimum units of nucleosomes, in which a histone core is surrounded by two winds of DNAs (140–50 bp). The histone core includes the pairs of H2A, H2B, H3, H4 histones whose molecular weight is about 10,000 Da each. A chromatin is formed from nucleosomes linked with about 50 bp DNA, and then a chromosome is made of chromatin folded by proteins such as the condensin. Chromatin had been considered as just a folder of the giant DNAs to pack into a nuclear whereas now we know that chromatin plays important roles to gene functions. Euchromatin is the domain involving an activated transcription, which has a loosen structure between nucleosomes. Heterochromatin is a domain of suppressed transcription, which has a highly condensed structure. Genomic regions with few genes, such as centromere and telomere, also have heterochromatins. The structural change between an euchromatin and heterochromatin is mediated by modifications in both DNA and histone, which is called the chromatin remodeling and regulates gene transcription (Muller and Leutz 2001; Nair and Kumar 2012).

10.1.4 DNA Modification

DNAs can be methylated on cytosine residues, and these modifications could be epigenetic and be inherited through a cell division. However, DNA methylations could be reprogramed during gametogenesis.

DNA methylations take an important function role in both development and differentiation of a body, and those are inherited from old DNAs to new DNAs upon replication. DNA methylations mainly induce transcription repressions. Some of DNA methylations could take place reversibly depending on an external factor as environmental change. The DNA methyltransferases (DNMT), DNMT1-3 found in human are the enzymes which add methyl groups on DNAs. It had not been clarified until recent which enzymes are involved in the DNA demethylation but Tet proteins have been found in 2004 which could hydroxylate a methylated group (Tahiliani et al. 2009).

DNA methylations can be classified into two functional types, maintenance or *de Novo*. Maintenance methylations are observed during a DNA replication that copies methylations of an old DNA to a new DNA. *de Novo* methylations occur after reprograming. DNMT1 has been considered to be responsible for maintenance methylations since null-mouse of DNMT1 is embryonic lethal. It has been confirmed by our research group that DNMT1 indeed converts hemi-methylated DNAs to fully-methylated DNAs by recognizing histone H3K23 ubiquitin mark added by uhrf1 (Nishiyama et al. 2013).

DNMT3s are also considered to be involved in *de novo* methylations, and their three subtypes (3a, b, L, 3a and b) are involved in DNA methylations during early development. Among them, L has no enzymatic activities. Although DNMT2s are categorized as DNMTs by their sequence homologies, DNMT2s have methyltransferase activities not for DNAs but for tRNAs (Szyf and Detich 2001; Turek-Plewa and Jagodzinski 2005).

10.1.5 Histone Modification and Variants

Histone acetylations had been revealed after discovery of DNA methylations. The interaction between DNAs after histone acetylations is weaken by their nature of basicity, and then the chromatin turns to euchromatin which activate transcription. Histone acetyltransferases (HATs) are the enzymes that acetylate histone, and reversely histone deacetylases (HDACs) deacetylate histones. Both DNA methyltransferases and histone acetyltransferases have been the therapeutic targets for anti-cancer drugs. Some are available commercially on market, including the HDAC inhibitors, Zolinza and ISTODX, for the cutaneous T cell lymphoma (CTCL) and the DNMT inhibitors, VIDAZA and DACOGEN, for the myelodysplastic syndromes (MDSs). It is not only acetylation, but also other modifications including phosphorylation, methylation, ubiquitination, SUMOylation, and citrullination that to mediate epigenetic regulations. In recent decades there have been other modifications found such as crotonylation, propyonilation, butyrylation for lysine, acetylation for serine/threonine and succinylation for lysine. All of the modifications have not been proved to be involved in epigenetic regulations yet, but their functional insights would be understood in near future. Researches on histone methylations have been proceeding rapidly in parallel to the study

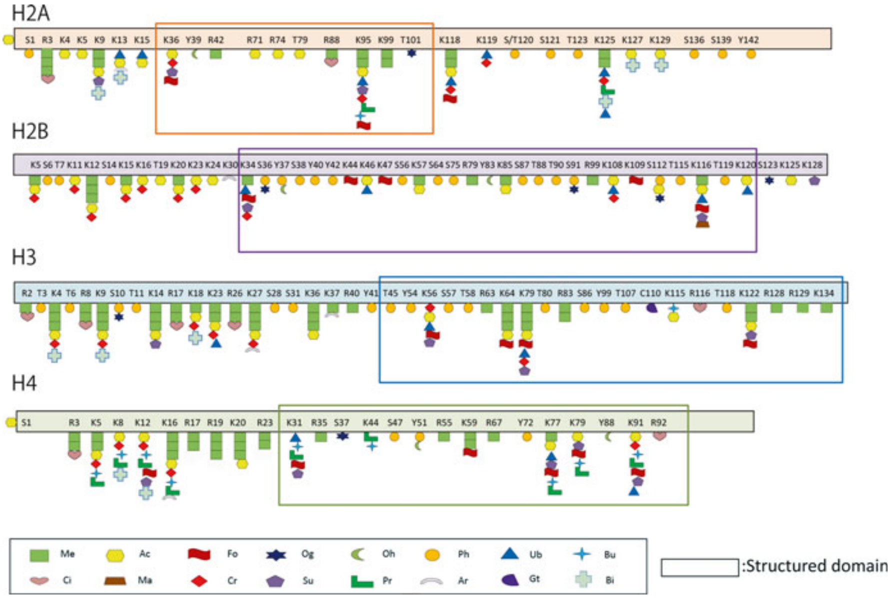

H2A
S1 R3 K4 K5 K9 K13 K15 K36 Y39 R42 R71 R74 T79 R88 K95 K99 T101 K118 K119 S/T120 S121 T123 K125 K127 K129 S136 S139 Y142
H2B
K5 S6 T7 K11 K12 S14 K15 K16 T19 K20 K23 K24 K30 K34 S36 Y37 S38 Y40 Y42 K44 K46 K47 S56 K57 S64 S75 R79 Y83 K85 S87 T88 T90 S91 R99 K108 K109 S112 T115 K116 T119 K120 S123 K125 K128
H3
R2 T3 K4 T6 R8 K9 S10 T11 K14 R17 K18 K23 R26 K27 S28 S31 K36 K37 R40 Y41 T45 Y54 K56 S57 T58 R63 K64 K79 T80 R83 S86 Y99 T107 C110 K115 R116 T118 K122 R128 R129 K134
H4
S1 R3 K5 K8 K12 K16 R17 R19 K20 R23 K31 R35 S37 K44 S47 Y51 R55 K59 R67 Y72 K77 K79 Y88 K91 R92
Me
Ac
Fo
Og
Oh
Ph
Ub
Bu
Ci
Ma
Cr
Su
Pr
Ar
Gt
Bi
:Structured domain

Fig. 10.2 Latest core histone modifications (*Ac* acetylation, *Fo* formylation, *Og* O-glcNAcylation, *Oh* hydroxylation, *Ph* phosphonylation, *Ub* ubiquitination, *Bu* butyrylation, *Ci* citrullination, *Ma* malonylation, *Cr* crotonylation, *Su* succinylation, *Pr* propionylation, *Ar* ADP-ribosylation, *Gt* glutathionylation, *Bi* Biotinylation) (Arnaudo and Garcia 2013; Graff and Tsai 2013; Tweedie-Cullen et al. 2012; Zempleni et al. 2008) HIstome: http://www.actrec.gov.in/histome/ etc

progresses on acetylations (Arnaudo and Garcia 2013; Graff and Tsai 2013; Tweedie-Cullen et al. 2012) (Fig. 10.2).

Since a histone methyltransferase had been found first in 2000, about 60 kinds of methyltransferases have been reported. The demethylase, LSD1, was found at first in 2004 and about 40 kinds of demethylases have been counted since then. A histone methylation has little charge effect unlike acetylation, and therefore it has been considered to have an effect on interactions between histones and other proteins. Whereas acetylations occur on lysines, both lysines and arginines can be methylated. In fact, the state of histone methylation affects gene expression since methylations induce an activation of transcription in the case of histone H3K4 although the K9 and K27 methylations could inactivate transcription. The recent studies have demonstrated that these modifications would result in different effects when other modifications are combined: i.g. in the co-existing case of both the activation mark (H3K4me3) and the repression mark (H3K27me3) (Kooistra and Helin 2012), which corresponds to the "waiting mode" so that the demethylation of H3K27me3 leads to an immediate activation of transcription.

Histone modifications are observed mainly on a histone tail, which is N-terminal of each histone and mostly unstructured. Rests of a histone (C-terminal) consist of well-structured cores and wrapped around with DNAs. The longest histone tail is of histone H3, about 50 amino acids. Histone tails have numerous modifiable lysines and arginines, and so combinations of acetylation and methylation count theoretically up above 200 kinds for H4 and 20,000 for H3. Taking into account all of the known modifications, the number of possible combinations goes up above a trillion for H3. Novel modifications would add more. These combinatorial variations of modifications in several tens of amino acids would determine gene expression to guide to form individually differentiated cells of ca. 200 kinds which make a human body.

These modifications can be reversible. The following brief names describing the processes of modifications are given. An enzyme that adds a modification is referred as "writer", that erases is "eraser", and a protein recognizing a modification is "reader". "Writers" include DNMT, HAT, and HMT (histone methyltransferases). "Erasers" include HDAC and KDM (lysine demethylases). "Readers" can be classified into several functional modes: (1) Binding modifications such as the Bromo domains for lysine acetylations, (2) MBT, CHROMO and PHD domains for methylated lysines. (3) PWWP domains for methylated lysine such as H4K20me that relate to preservation/prevention from DNA damages and their transcription, (4) TUDOR domains for RNA (Table 10.1).

Histones are regulated not only by their modifications but also by their variants. Histones form an octamer of two pairs of heterodimer of H2A and H2B and that of H3 and H4. During a nucleosome formation, DNAs first bind to the complex of H3 and H4, followed by the complex formation with H2A and H2B. H4 has a highly conserved sequence and have no variants. The most abundant H3 is H3.1, but other than that, there are H3.2, H3.3, H3t and CENP-A. H2A includes H2AX, H2AZ, H2ABbd and macroH2A, and H2B includes spH2B, H2BFWT, nTSH2B (H2A and H2B have other small difference variant).

Among human H3 variants, H3.1 and H3.2 are incorporated into a chromatin when DNAs are replicated. H3.3 is incorporated in chromatin where transcription is active (Fig. 10.1). CENP-As have specific functions and are located only on pericentromere regions, that CENP (centromere protein) is named after, and their homology shares only about 50 % while other variants are different only a few amino acids. H3t is a variant expressed in testis but its function remains unknown.

Among the H2A variants, H2A.X functions are in repairing damaged DNAs and recruiting DNA repair proteins. H2A.Z activates and suppresses of transcription, which seems to be a highly related to carcinogenesis because its up-regulation has been reported frequently in various cancers. It might be involved in cell proliferation and genomic instability. However, a H2B variant still remains lacking and unknown. Thus, numerous variants of histones could drive to regulate chromatin structures and transcription similarly to modifications (http://www.actrec.gov.in/histome/).

10.1.6 RNA

RNAs being transcribed from DNAs are also one of epigenetic players. The classical central dogma showed a straight information flow from DNAs (Genome) ->RNAs (Transcriptome) ->Proteins (Proteome). mRNAs transcribe from genes to make up

Table 10.1 Epigenetic modification-related proteins

WRITER					ERASER	
HMT			DNMT	HAT	HDAC	KDM
SETD7	SETD6	PRDM1	DNMT1	ELP3	SIRT5	KDM3B/JMJD1B
SETD8	SETD4	PRDM14	DNMT3B	KAT2B/PCAF	SIRT7	KDM3A/JMJD1A
EZH1	SETD3	PRDM12	DNMT3A	KAT2A/GCN5L2	SIRT6	JMJD1C
EZH2	SMYD4	PRDM6	DNMT3L	MYST4	SIRT4	NO66/FLJ21802
MLL2	SMYD5	PRDM4	TRDMT1	MYST3	SIRT3	MINA
MLL3	SMYD1	PRDM10		MYST2	SIRT2	KDM2B/FBXL10
MLL	SMYD2	PRDM15		MYST1	SIRT1	KDM2A/FBXL11
MLL4	SMYD3	PRDM11		KAT5/TIP60	HDAC7	PHF8
SETD1A	SUV420H1	PRDM7		GTF3C4	HDAC9	JHDM1D/KIAA1718
SETD1B	SUV420H2	PRDM9		HAT1	HDAC5	KD 1B/LSD2
SETDB2	CARM1	PRDM2		TAF1L	HDAC4	KFM1A/LSD1
EHMT1/GLP	PRMT2	PRDM16		TAF1	HDAC10	KDM6B/JMJD3
EHMT2/G9q	PRMT6	MDS1		ATAT1	HDAC6	UTY
SETDB1	PRMT1	PRDM5		EP300	HDAC11	KDM6A/UTX
SETMAR	PRMT8	PRDM13		CREBBP	HDAC8	KDM5D/JARID1D
SUV39H1	PRMT3	PRDM8		CLOCK	HDAC3	KDM5C/JARID1C
SUV39H2	PRMT7			NCOA1	HDAC2	KDM5B/JARID1B
MLL5	PRMT5			NCOA3	HDAC1	KDM5A/JARID1A
SETD5	DOT1L					KDM4DL/JMJD2E
O6ZW69						KDM4D/JMJD2D
ASH1L						KDM4C/JMJD2C
SETD2						KDM4B/JMJD2B
WHSC1L1						KDM4A/JMJD1A
NSD1						JARID2
WHSC1						

http://www.sgc.utoronto.ca/epigenetics/domain_tree proteins and to form a life by the function of tRNAs. Non-coding RNAs (ncRNAs), that does not code proteins, would play an important function in epigenetic regulations. ncRNAs include small interfering RNAs(siRNA) and microRNAs (miRNA). These are short RNA fragments complementary to mRNAs, which interact with mRNAs to abolish protein synthesis. It is estimated that ncRNAs consist of total 40 % of the whole genome, that should be compared to the transcribed regions of 2 %. ncRNAs have been increasingly recognized on their importance although it was previously thought as just a junc. Thus, DNAs, RNAs and proteins co-operate closely for epigenetic regulations in development, growth and maintenance of a body.

RNA*i*s (inheritance) break mRNAs which is complementary sequences of two-strand RNAs. This phenomenon is initially considered as a response against external RNA viruses. Two-strand RNAs that cause RNA*i*s are called *si*RNA. Two-strand RNAs has been demonstrating an abolishment of gene expression more effectively than antisenses that have the similar effect. An RNA*i* is produced via cleavage of a long RNA by a nuclease Dicer, *si*RNA, and then it forms a complex with other nucleases (Argonaute families) to form a RISC (RNA-induced silencing) complex. This RISC complex degrades the target RNA to suppress translation. This mechanism can be applied artificially to suppress genes, in which the *sh*RNA of hairpin structure is incorporated in an expression vector. When the vector is transfected to cells, some *sh*RNAs involve the production of the interferon γ, which implies that RNA*i*s could be originated from an anti-virus system.

*mi*RNAs are of the similar character as *si*RNA. A *si*RNA is consists normally of 21 base pairs long and a miRNA is similarly of 20–25 bases long. Dicer cleaves out a pre-*mi*RNA into a mature *mi*RNA . However, *si*RNA breaks mRNA whereas *mi*RNA does not break mRNA and inhibit translation. Recently it has been reported that a protein forming a complex with a *mi*RNA is recruited to the complimentary RNA near elongating protein, and that the complex then turns an adjacent chromatin to a heterochromatin. *mi*RNA can be one of the drug targets, and antisenses targeting miRNAs have been developed to cure cancers (Di Leva and Croce 2013; Singh et al. 2013).

10.1.7 Polycomb and Trithorax

Researches in epigenetic regulation had been advanced in plants, and it has been noticed that Drosophila has the similar system. Important genes in early development in animals include Hox genes. Hox genes are involved in body segmentation and highly conserved among animals. In Drosophila, it determines the segmentation of head, thorax, and abdomen, and in human, it determines growth of legs and other segments during fetal development.

The polycomb complex is the protein complex that is involved in the expression regulation of Hox genes. It was reported over 30 years ago that this complex causes an abnormality of segmentation in Drosophila (Denell 1973). In this complex, the histone methyltransferase, EZH2, is involved, which was first cloned as the factor

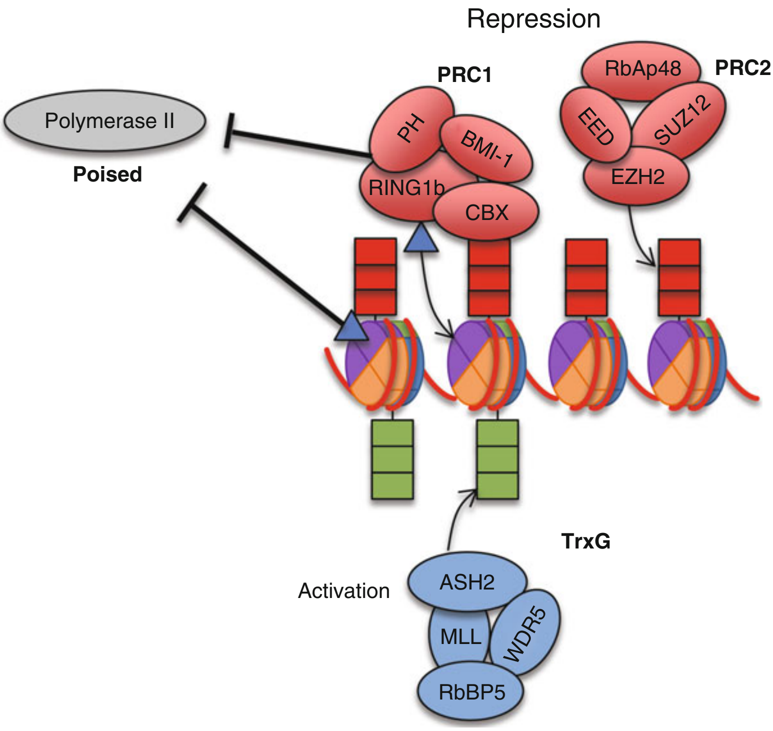


Fig. 10.3 Polycomb and trithorax (Richly et al. 2011)

that determines transcription and chromatin structures, and it was found in 2002 that this complex has the SET domain that methylates histones (Kuzmichev et al. 2002).

EZH2, as well as SUZ12, EED, and AREBP, participates in the PRC (Polycomb repressive complex) 2, and has the activity to mediate H3K27 to mono, di, and tri-methylations. The somatic mutation Y641F/N activates tri-methylations that would cause the diffuse large B-cell lymphoma (DLBCL). The PRC2 complex methylates H3K27 in the region that codes Hox gene, where methylated H3K27 is bound by PRC1 to induce inactivation (Fig. 10.3). In contrast, a trithorax complex (TrxG) is the counterpart of the PRC2 and activates the same gene. TrxG is also a complex containing histone methyltransferases including MLL, and methylates H3K4. The methylated H3K4 recruits a chromatin-remodeling factor to activate transcription. In Drosophila, the PRC complex determines abdomen whereas TrxG complex determines thorax and their appropriate balance leads normal segmentation. However, it is not clear whether they are indispensable factors. These methylations of H3K4 and H3K27 are known as epigenetic marks that regulate transcription not only during development. For example, H3K4me3 is located widely on promoter regions of activated genes and H3K4me2 is found in the latter part of genes. The transcription repression mark, H3K27me3, is detected on near the replication origin (Fig. 10.1).

EZH2 in the PRC2 complex is overexpressed in many types of solid cancers, and have a positive correlation with progression of cancers such as breast, prostate, non-small cell lung cancers. Therefore, EZH2 is one of attracting drug targets, for which big pharmas like GSK and Esai (epizyme) are indeed developing. Recent researches for hormone-independent prostate cancer have shown that EZH2 can have an oncogenic activity independent from the PRC2 complex, implying that a histone methyltransferase can mark differently on gene to gene by forming a distinct complex (Xu et al. 2012).

10.1.8 Xist and Heterochromatin Formation

As described previously, chromatins are highly condensed into a Barr body during X-inactivation. HP1 (heterochromatin protein 1) is a protein that is located on heterochromatin. The proteomic studies together with next generation sequencing by using HP1 as a bait suggested the mechanism in which a Barr body is formed by a long highly-specific *nc*RNA XIST(X-inactive specific transcript), H3K9me3, H3K27me3 and HP1 marks (Nozawa et al. 2013).

Dosage compensation is taken place by X-inactivation in mammals. However, it could be achieved by reducing transcription by half in female as in C. elegance or increasing transcription twice in male as in Drosophila. Transcription must be in any case the same between male and female. When such the balance is disrupted, it can cause inborn disorder or embryonic lethal. As *nc*RNA XIST is involved in X-inactivation in human, X-inactivation in Drosophila is also mediated by *nc*RNA. In Drosophila, two kinds of *nc*RNA roX1 and roX2 bind to a male X-chromosome to double transcription. In this way, *nc*RNA can even increase transcription by changing a chromatin structure, or reduce transcription by forming the complementary strand.

10.1.9 Reprogramming

Epigenetic modifications including DNA methylation, histone modification, non-coding RNA can be erased upon a specific timing, which is referred as "reprogramming".

"Reprogramming" seemed to occur at the two stages, during gametogenesis of sperm and ovum, and between fertilization and early development. Zygote gradually loses their versatility to finally become one of differentiated 200 cells that consist each organ. "Reprogramming" corresponds to a removal of all of epigenetic modifications in DNA and histone. Recent studies have revealed that some of the modifications are stored among generations.

An interpretation for the incomplete reprograming can be given as follows: On the gametogenesis, DNA methylation is once reset during gamete differentiation, and re-written before gematogenesis after miosis while histone modifications are partially restored. Sperms have more condensed chromatins by substituting histones to protamines, but they still have histones with modifications. These remained modifications on histones could mediate the incomplete reprograming.

In the reprogramming after fertilization, genes without a genomic imprinting are reprogrammed to differentiate into a variety of cells.

This storage allows the inheritance of a trait without any gene mutation, which cannot be definitive epigenetic inheritance in a limited sense, but by which unaccountable phenomena by Mendel's rule might be explained. The theory of inheritable acquired traits that was given by Lamarck does not seem to be denied completely by the mutation theory of Mendel. Among geneticists, there have been some unaccountable phenomena to which rigorous answers have not been given for a long time but now epigenetic approaches might be ready to answer.

Epigenetic inheritance is now uncovered, exampling such that habits of mother in her pregnancy affects her child constitutions and that drugs exposed to father induces abnormal spermatogenesis without DNA mutation that is indelible even in grandchild (Anway et al. 2005; Anway and Skinner 2008). Epigenetic marks are the modifications that cause epigenetic inheritances, and it is known in plant that epigenetic marks are robust to be maintained through generations.

Although there are symmetric and asymmetric dimethylations on arginines, those seem to be distinguished in reprogramming. Reprogramings take place on gametes, zygotes and somatic cells, where asymmetric dimethylations are observed mostly in germ cells while symmetric ones are in somatic cells.

iPS cells are generated by forcing reprogramming. While a zygote is differentiated into a body, cells gradually lose their differentiation potencies: totipotent (completely versatile) zygote, pluripotent embryonic stem cells (versatile but for placenta), and through to multipotent (limited potency) somatic stem cells, including hematopoietic stem cells and neural stem cells. Somatic stem cells include cancer stem cells (CSCs), which lead the recurrence of carcinogenesis. Then, those cells become oligopotent cells, which are more reduced in the direction of development and finally become one of unipotent cells of about 200 kinds with an unpotency that form a specific each organ.

Stem cells are characterized by its asymmetric cell division. Unipotent cells are divided into two identical cells while stem cells are divided into one stem cell and one differentiated cell. Differentiated cells then continue to divide and proliferate, while the number of the stem cells does not change. It should be emphasized that iPS cells have an artificial pluriopotency acquired by reversing this differentiation process (Fig. 10.4).

As an example of epigenetic inheritance over generations, in human, it is known that fat fathers tend to have underweight babies. Though it is not easy to distinguish genetic or epigenetic, an experiment with rats proved this as epigenetic. Fathers given high fat food had underweight babies and the reduced insulin release evoked by glucose to have abnormalities in sugar resistance. These rats had a reduced number of islets of Langerhans in their pancreas and an increased expression of the interleukin-13 receptor α2 (IL-13α2). When methylations were investigated around

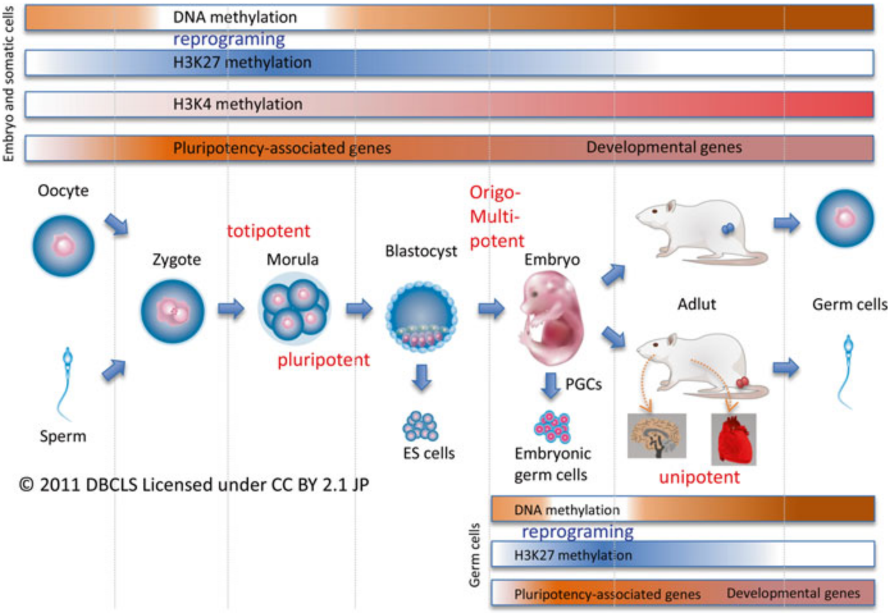


Fig. 10.4 Epigenetic regulation during mammalian development (Reik 2007) etc

this gene, methylations were found to be reduced, indicating that a habit of father to take high fat food seems to cause a hypomethylation of the gene coding IL-13α2 and this epigenetic change most probably inherited to his daughter. Although it is still unknown how the epigenetic change escaped from reprogramming to be reserved, this example has very interestingly shown that the habit of fathers not mothers can affect his daughter (Ng et al. 2010).

The above example showed a true epigenetic inheritance, but the effect of habits of mother on baby in her pregnancy is not an entirely epigenetic. That should be correctly understood by that the reconstruction of epigenetic marks after reprogramming is affected by environment, which could result in different phenotypes in monozygotic twins.

10.1.10 Cell Division

Both histone modifications and DNA methylations maintain chromatin structures, and the higher order structure of a chromosome is formed by the following additional proteins. Among them, cohesin regulates the adhesion of sister chromatids, condensin is involved in chromosome segregation, and CTCF as a insulator.

In mitosis, cohesions adhere replicating DNAs during S-phase so that synthesized DNAs are prevented to be separated. Then, in M-phase it is replaced to condensin to allow a chromosome to be segregated more densely to be sister chromatids.

Sister chromatids make a pair with the two identical chromosomes to form a bivalent chromosome, where cohesin forms a different complex as in mitosis. Then this bivalent chromosome is separated into identical chromosomes in the first division, and sister chromatid in the second division. If cohesin complex has abnormality, separation of chromosome goes wrong to induce several diseases. Such a structural change in whole chromosomes is under epigenetic regulation. Mutations in cohesin complex genes are frequently recognized in myelodysplastic syndromes (MDS), Chronic Myelomonocytic Leukemia (CMML), Acute myelogenous leukemia (AML) and Chronic myelogenous leukemia (CML) (Kon et al. 2013).

10.1.11 Retrotransposon

Genes coded on human genome consist only 2 % of the genome and above half are repeating sequences. Most of the repeating sequences are the transposable sequences called transposon. Retorotransposons are once transcribed to RNA and reverse transcribed to DNA to be inserted in genome while the transposons move directly as DNA.

Retrotransposons are classified into LTR (long terminal repeat) and non-LTR by their repeats. Non-LTRs are further sub-classified into the long interspersed nuclear element (LINE) and short interspersed nuclear element (SINE) by the length of the repeat sequences. LINEs consist 21 % and SINEs 13.5 % of the human genome. LINEs resemble to mRNAs whereas SINEs to tRNA and rRNA. Retrotransposons are normally highly methylated so as to suppress transcription.

10.1.12 None-Histone Target of Epigenetic Enzyme

Others than histones can be acetylated. Most protein N-terminals can be acetylated but lysine acetylations seem to relate pathogenesis. Tumor suppressor p53 is also acetylated. Acetylation of p53 allows p53 to bind DNA through its DNA binding domain to activate transcription of genes that inhibit proliferation and promote apoptosis. Except for the function of transcription factors, p53 have been shown to induce microRNA to suppress carcinogenesis (Dai and Gu 2010). Thus, many of transcription factors and co-factors are involved in epigenetic regulations.

10.1.13 Vitamin

Recently, it is recommended to take folic acid during pregnant to prevent congenital abnormality, as in the United States, where folic acid is added in foods and spread in market as supplement. It is now revealing that folic acid is essential in DNA synthesis and, as other than that function, it has ability to induce epigenetic changes during development. Also, another familiar supplement in US SAM (S-adenosyl methionine) to prevent depression is also a co-factor for epigenetic enzymes. SAM donates methyl group when DNMT, HMT mediates methylation. When SAM is depleted, methylation state of the body goes down. Ancient Chinese (herb) medicine and folk medicine have effects in such way could including epigenetic factors.

10.1.14 Infection

Besides cancer and stress, infection is also reported to accompany with epigenetic regulations. Whereas it is well known that cancers have highly methylated DNAs, the infection by the Helicobacter pylori, a mediator of stomach cancer, induces epigenetic changes in gastric mucosa. A comparison on DNA methylations between infected individual and normal has showed the elevated level of methylations in the region of stomach cancer and that it was reversed by the pylorus eradication (Ando et al. 2009; Maekita et al. 2006). Although it is unclear whether virus actively induces epigenetic changes or the changes were passively induced under the stress of infection, the epigenetic regulation certainly mediates there. Some viruses infect by suppressing epigenetic regulations in a host. Influenza viruses have the histone mimetic NS-1 that interacts with transcription factors instead of H3 tail to suppress anti-virus responses (Marazzi et al. 2012; Tan et al. 2013). So NS-1 could be a kind of epigenetic inhibitors.

10.2 Approach for Epigenetic Analysis

10.2.1 Introduction

In the previous section, we have described about epigenetic regulation. Next, we introduce the methodologies for epigenomic analysis and drug discovery. As described previously, epigenome is all about targeting DNA methylation, histone modification, non-coding RNA (ncRNA), histone modification recognizing protein complexes. To analyze those complex objects, the recently advanced technologies of next-generation sequencing and mass spectrometry (MS) have been demonstrating their powerful capabilities to explore a frontier of epigenomics. Herein, those comprehensive analyses and their derived whole information are defined as epigenomic analysis and epigenome, respectively.

10.2.2 Epigenetic Regulators as Drug Targets

Recently, the relationship between epigenetic regulation and disease is highlighted and attracting much attention next to protein kinases. In post genome sequencing era, the main stream to discover drug target have been based on genetic mutation analysis. Biomarker discovery against mutations such in tumor suppressor p53 and receptor tyrosine kinase EGFR was strenuously conducted. Also, therapy strategy has shifted from global drug treatment to personal medicine, where the mutation analyses have been justified. However, contrary to expectation, not so much novel biomarkers have been discovered. In such a situation, it was found that methylation in CpG islands through whole genome decreases but methylation on promoter regions of a tumor suppressor gene is up-regulated. Demethylation at CpG islands would cause the instability of house-keeping genes while the methylation of promoter of a tumor suppressor gene suppresses its expression (Esteller 2008). Moreover, changes in the state of histone methylation were found in various types of cancer cells (Mohammed et al. 2012; Rodriguez-Paredes and Esteller 2011; Waldmann and Schneider 2013).

Thus, the epigenetic mark was found to vary in diseases without any genetic mutation. Since we cannot recover genetic mutations, we had focused on killing cancer cells in the cancer treatment so far. In epigenetic diseases, we could recover the changes in epigenome using drugs. As drugs targeting epigenome has already been introduced in market, which include HDAC inhibitors and DNMT inhibitors. Histone methylations are attracting attentions. So far, it has been found that dozen of HDAC and three kinds of DNMT but about 60 and 40 kinds for HMT and KDM, respectively. This diversity might give a disease specific methylation pattern and at the same time it is expected that targeting one of these enzymes might exhibit a high specificity (Fig. 10.5, Table 10.1) (Table 10.2).

The combination of histone modifications is called as the histone code, and this histone code *per se* is expected as a biomarker. For example, the modifications in histone H4 are associated with malignant transformation. For H3, the Polycomb complex protein EZH2, that methylates H3K27, are overexpressed in various cancers and SMYD3, that methylates H3K4, are overexpressed in colon, liver and breast cancers. Most of the H3K27me3 and H3K9 methylations are associated with the suppression of transcription while most of the H3K4 methylation are associated with the activation of transcription, which suggesting that there are genes that are suppressed or activated by an epigenetic mark (Yost et al. 2011). Discovery of the code that activate tumor suppressor or suppress oncogene could lead to finding new therapeutic targets by drugs of new chemical entities or based on antibodies.

Currently such biomarkers are detected and quantified for their acetylation in blood or cancer (Wagner et al. 2010). Generally, hyperacetylation leads to a transcription activation while hypoacetylation does a chromatin condensation so as to suppress transcription. An examination of the acetylation state can be performed by measuring to assess effectiveness of treatment of HDAC inhibitor.

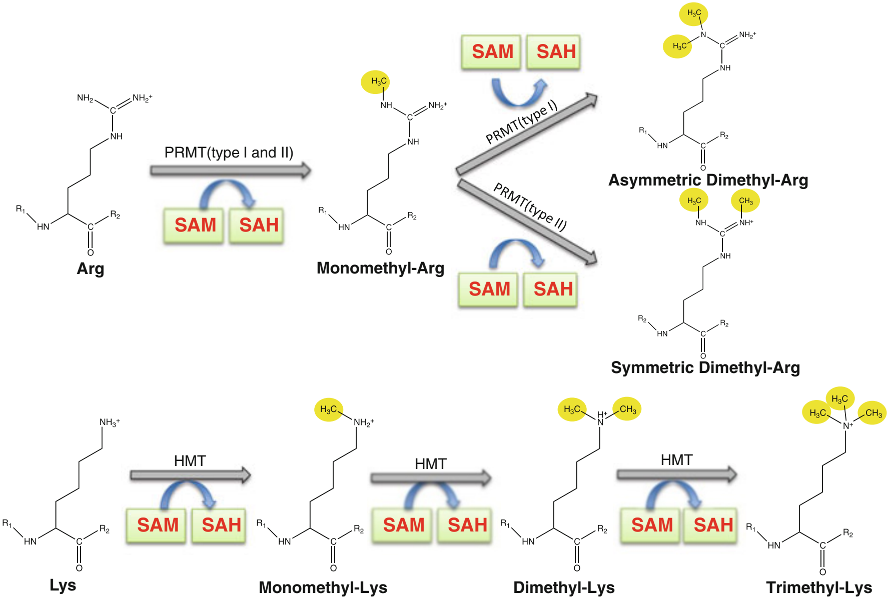


Fig. 10.5 Reaction of protein methylation

Table 10.2 Epigenetic-enzyme inhibitors for cancer therapy

Generic name	Alternative names/Code Nr.	Clinical status	Application
DNMT (DNA methyl transferase) inhibitors			
5-azacitidine	Vidaza	FDA approved April, 2004	Myelodysplatic syndrome (MDS)
Decitabine	Dacogen	FDA approved May, 2006	MDS
HDAC (histone deacetylase) inhibitors			
Vorinostat	Zolinza	FDA approved October, 2006	Cutaneous T cell lymphoma (CTCL)
Romidepsin	Isodax	FDA approved November, 2009	CTCL
Panobinostat	LBH-589	Phase III	Adult progressive solid cancer, CTCL
Belinostat	PXD-101	Phase II	Neoangiogenesis
Entinostat	MS-275, SNDX-275	Phase II	Vascular tumor
MGCD-0103	MG-0103	Phase II	Vascular tumor
JNJ-26481585	None	Phase II	Vascular tumor
Givinostat	ITF2357	Phase II	Vascular tumor, multiple myeloma

Copeland et al. (2009) with modifications

10.2.3 Epigenome Associated Protein Complex as Drug Target

The histone acetylation state might link mainly to a chromatin structure, and its methylation state might link to an interaction with their reader proteins. Depending on modifications, different reader proteins are recruited to modifications, and their protein complex could induce a chromatin remodeling so as to regulate gene expression. An interaction between these complexes and epigenetic modifications might also lead to drug targets.

To inhibit enzymes, generally, we target their active center as protein kinase inhibitors. However, this strategy is not effective to target complex formation. As mentioned above, EZH2s are active only when they form complex with EED and Suz12. Thus, to target EZH2 activity, we could interrupt not against EZH2 active center but this heterotrimer interfaces. We need a new strategy to conduct for this purpose. If we already had structural data about a target complex, we could design drugs based on its structure, if not, we hardly obtain hit compounds even by screening through 10^5 order of compounds. For this objective, the Random Peptide Integrated Discovery (RAPID) is available (Hipolito and Suga 2012) (http://www.peptidream.com/). The RAPID is a screening system in order to search candidate peptides which are interactive to the target protein by using a circular peptide library containing 10^12 repertories, which number indeed exceeds those of antibodies available. We have been so far discovered some of interactive peptides against some targets by this approach (unpublished).

10.2.4 Development of Epigenetic Regulator Inhibitor

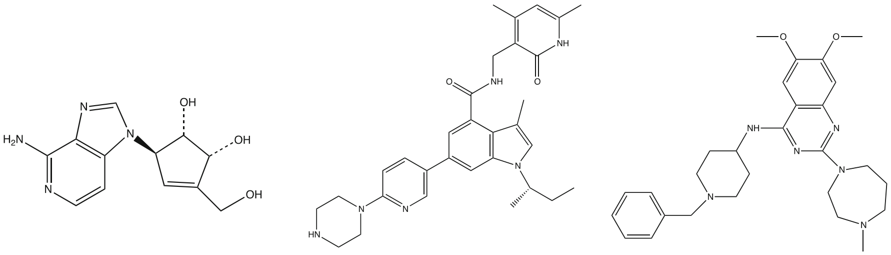
We introduce here some examples of drug development that targets histone methyltransferase. Histone methyltransferases have a motif called SET domain. SET was named from initial letters of the Drosophila genes SUVAR3-9, Enhance of zest and TRITHORAX that had a common domain and function in position effect variegation (PEV) for transcription suppression. Most of these proteins have activities that methylate lysine. SET domain works as an active center where substrate histone and methyl donating co-factor S-adenosylmethionine (SAM) bind there to mediate methyl transfer reaction (Fig. 10.5).

Most of the known HMT inhibitors aim to target this active domain. None of these HMT inhibitors are on market yet, but several drugs have been reported to include those targeting EZH2s. The EZH2 forms a PRC2 complex, which is active as a trimer with EED and Suz12 and mediates to form H3K27me3. The first reported HMT is a SAM homolog DZNep to inhibit most of HMT but have certain selectivity against EZH2. The DZNep inhibits the H3K27 methylation of the PRC2 complex selectively to induce apoptosis in cancer cells (Tan et al. 2007). When the SAM is converted into the SAH (S-Adenosylhomocysteine) by donating its methyl group, it acts as a pan-inhibitor for methyltransferases with a feedback inhibition.

There are other SAM analogs with similar effects. GSK126, which has been developed by GSK and reported in 2012, is a specific drug targeting EZH2 with the IC50 of several nM *in vitro* and (McCabe et al. 2012). This compound has been thought to inhibit the histone binding and is membrane permeable. GSK126 has shown to be competitive to not only histone but also SAM, and inhibit the proliferation of the DLBCL lymphoma cells that contains a mutant EZH2 (Y641) with an enhanced trimethylation activity. This compound inhibits the cancer growth in the DLBCL mouse model of bile cancer, indicating that EZH2 in EZH2 mutant cells can be a therapeutic target (Kipp et al. 2013). With GSK126, a number of papers have been reported on how PRC2 does function. Generally, in drug development, we first identify seed compounds based on their X-ray structures, and then leads are designed from seed compounds. A number of research teams have been tried to solve a 3-dimensional structure of the PRC2 complex because that is highly expected to result in discovery of epigenomic drugs. The structural genome consortium (SGC, http://www.thesgc.org) for epigenome drugs has been established for their systematic analysis. SGC and big pharmas have been cooperating together to conduct the structure-based drug screening.

Our group has been also involved in developing HMT inhibitors, including PR-SET7 inhibitor. So far, as in the discovery of DZNep and GSK126, they were found in their screening through numerous compound libraries stocked in each company. Whereas the number of screened compounds is unknown in reaching GSK126, Boehringer Ingelheim obtained BIX-01294 (G9a inhibitor) by screening through 125,000 compounds (Kubicek et al. 2007). Globally at present ten-million compounds are available for us but it does not seem realistic to screen all of those compounds (Fig. 10.6).

To increase efficiency, the *in silico* screening has been conducted by using supercomputing against three million of compounds with molecular weight around 300 Da that are suitable for synthetic optimization. By rapid increase in CPU power, we are now able to calculate molecular dynamics between interacting proteins (Martin Karplus, Michael Levitt and Arieh Warshel 2013 Noble prize). We first identify compounds that can dock into a substrate pocket based on the X-ray structure of a target protein. Screening of dockings into HMT is carried out by avoiding



3-deazaneplanocin A
(DZNep)

GSK126

BIX-01294

Fig. 10.6 Compounds of HMT (EZH2, G9a) inhibitors

the SAM binding region to prevent from pan-inhibitors. Then, hit compounds are narrowed down by examining their actual reactions. In such regressive repeats of *in silico* and *in vitro* screening, molecular dynamical structures of both a docking compound and an enzyme are calculated to investigate their stabilities and their synthetic optimizations. Thus, the inhibitor EBI099 has been attained against PR-SET7 and HMT that regulate cell cycle progression. This compound exhibited its stable docking into the PR-SET7 according to its molecular dynamical calculations, and showed somewhat high IC_{50}. Its structural optimization would provide 2-order of magnitude high in IC_{50} by affinity enhancement. Numerous inhibitors obtained similarly against other HMTs have demonstrated that our approach to epigenomic drug discovery is quite promising.

10.2.5 Analysis with Next Generation DNA Sequencer (NGS)

The Next Generation DNA Sequencer (NGS) can be applied to determine targets at first step in epigenetic drug discovery. Most of the recent progresses described above were obtained by using NGS. In contrast to the classical Sanger method where long DNA is sequenced, NGS sequences massive short fragments followed by their reconstruction based on data processing. The latest NGS can sequence 600Gbp, which is 200 times of human genome, and the cost is reducing year by year. DNA methylations are indirectly detectable by the bisulfate sequencing method where an unmethylated cytosine is first transformed into an uracil and is then sequenced. Thus the comprehensive genome-wide analysis of DNA methylations is now available, by which numerous DNA methylations have been found in cancerous cells.

NGS also provides a comprehensive RNA sequence in place of a microarray. Such the RNA-seq can amplify *nc*RNAs without poly-A signal so that novel transcripts are detectable.

NGS also can be applicable to analyze relationships between histone modifications and DNAs. For example, since EZH2 methylates H3K27, H3K27 marked histones is enriched followed by sequencing the DNA in order to investigate how the genes are regulated by EZH2. This method is the Chromatin immune precipitation sequencing (ChIP-seq) that was developed in 2007 and can be applied to regions of specific sequences (Fig. 10.7) (Fujita and Fujii 2013; Rivera and Ren 2013). In the ChIP-seq method, after cross-linking chromatin with fixation such as formaldehyde, DNAs are enriched by the immune-precipitation of proteins or histone modifications with a specific antibody. Then, DNAs are sequenced to know the site on chromosome. Accumulation of the data using multiple antibodies reveals the relationships between histone modifications and DNAs.

The International Human Epigenome Consortium (IHEC, http://www.ihec-epigenomes.org) was established with the high expectation for discovering novel drug targets in 2010, and aims by utilizing the ChIP-seq method to sequence at least 1,000 epigenomes in cells closely related with human health and disease.

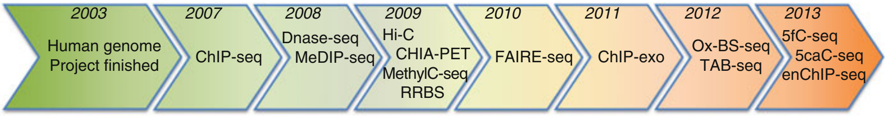


Fig. 10.7 Development of sequencing-based technologies (Rivera and Ren 2013)

DNA methylations are analyzed not only by the ChIP-seq method but also by MS-based sequencing. MS-based sequencing had been originally developed to detect SNPs, where a mass change is detectable when a SNP exist in the DNA sequence amplified by designed primers. The MS-based sequencing makes it possible for methylated fragments possible to be detected by combining with the bisulfate method. It is clinically applicable to utilize for diagnosis by detecting abnormal methylations, and so might be a novel diagnostic strategy (Jurinke et al. 2005) (http://www.sequenom.com).

10.2.6 Analysis with Mass Spectrometry

A recent advancement in MS-based proteomics has been highly progressive, not to the extent of NGS, and mass spectrometry is now indispensable technology in life science. The draft readout of Human Genome Project had completed in 2002, and then this rich information resource of human genome had been brought about into human proteome. Identification of peptides and/or proteins has been benefitted from the genome sequence readouts and has been revolutionized by recent advanced developments of mass spectrometry although only the time-consuming and laborious peptide sequencer or *de novo* sequencing were available until then. Advancements of mass spectrometry might make it possible to achieve a huge increase in the numbers of peptides and/or proteins identified. A current MS instrumentation has been providing an extremely high performance in sensitivity, scanning speed and mass resolution. It might not be far from reality that one measurement could identify and quantify the whole expressed proteins. Then, biomarkers might be more likely discovered from only several cells now. Proteome bioinformatics would become highly important to obtain a correct list of proteins identified and quantified from big proteomic datasets acquired on a MS, and to handle a complex MS/MS spectrum that does not simply match with genome sequence data. The similar problematic situation has been encountered in NGS regarding to BigData computing which is currently the major issue in life science.

Epigenomic analysis is performed by MS-based approaches. Reader proteins form a complex to bind the specific histone or DNA sequence. MS-based approaches are indispensable in the complex analysis because the direct determination of modifications can be achieved by using MS-based sequencing although indirect analysis of histone modifications regulating epigenetics is performed by ChIP-seq.

Whereas it has been known that specific modifications are relevant to specific diseases, e.g. H3K27me3 to cancer and H3K9 methylation to adiposity, the entire mechanisms of pathogenesis cannot be interpreted sufficiently by modifications but their various combinations. For example, chromatin regions with both repressive H3K27me3 and active H3K4me3 might be ready to transcribe (Standby mode) since the release of H3K27me3 immediately triggers transcription. The binary switch in the same histone or trans histone code in adjusting histones might regulate transcription (Fischle et al. 2003). Such a crosstalk could participate in a key role in epigenetic regulation. Some of the reader proteins can recognize a cassette of modifications but not individual modifications. To analyze these modification cassettes, a comprehensive analysis of histone modifications could be essential. Nextly, we describe MS-based analysis of epigenome, which differs from that of proteomics in several following aspects:

10.2.6.1 Chemistry of Histone

Most of histon modifications are observed in the N-terminal histone tail that is unstructured (Fig. 10.2). Histone is a basic protein and to form nucleosome by making a complex with acidic DNA. Histone is the molecule rich in lysine and arginine, both of which are modified frequently. A decrease in basicity of a histone derived by acetylation allows chromatin to form euchromatin to resulting in transcription activation. An acetylated lysine is also a target of reader proteins with the Bromo domain. A methylated lysine would have the function that recruits different reader proteins and ncRNA. The methylated arginine could have a function mainly in reprogramming which is its own way different from lysine. Arginine has two amino groups in the side chain, and so a dimethylated arginine can take two structures of symmetric and asymmetric (Fig. 10.7). These di-methylations are mediated by different enzymes and might have different functions.

10.2.6.2 Sample Preparation

Trypsin, in proteomics, is mostly used to digest proteins at C-terminal of lysine and arginine, and so trypsin digest histone into peptides of a few amino acids, which cannot be retained with any available LC columns. Moreover, in order to investigate the combination of modifications in the whole histone tail, not individual modifications, we need to cleave the intact tail. Most useful method to analyze histone modifications might be the lysine propionylation method. This method has been developed by Garcia group (Peters et al. 2003; Young et al. 2010; Young et al. 2009), and in which lysine is first propionylated to prevent from tryptic digestion, and then digested only at arginine. Therein, the combination of modifications cannot be known whereas individual modifications can be analyzed. Therefore, the approach based on top down proteomics without digestion or middle down approach with digestion of a preserved tail is necessary (Kalli et al. 2013; Yates and Kelleher 2013; Young et al. 2010).

10.2.6.3 High Performance Liquid Chromatography (HPLC)

Separation technologies for peptides dependent on their modifications are needed to analyze histone tail modifications. Whereas a conventional proteomic analysis utilizes a HPLC equipped with C18 reverse phase column (RP-HPLC). However, the acetylation of tryptic peptides will has their higher hydrophobicities, and so their elution times will be delayed whereas methylation has little effect on retention time. The Hydrophilic Interaction Chromatography (HILIC) is currently utilized to improve such the separation for peptides of hydrophilic histone tails by its strong retaining capability (Young et al. 2010).

10.2.6.4 Mass Spectrometry (MS)

Among major modifications, it is quite frequent that methylation takes place at multiple sites: e.g. lysine methylation can be seen at 1–3 sites, and arginine methylation commonly at 1–2 sites. It had been quite difficult in mass spectrometric analysis to distinguish between acetylation and trimethylation, whose mass difference is only 0.036 Da. This value is equivalent to 18 ppm in H4 tail (1-23aa) and 7 ppm in H3 tail (1-43aa). However, recent advances of high-accuracy mass spectrometry have made it possible to distinguish those with extremely high resolution less than 5-ppm. To analyze mixtures in a sample, the mass resolutions up to 200,000 for H4 tail and 500,000 for H3 tail are required. The FT-ICR MS can be operated under such high mass resolutions but its disadvantage is that longer acquisition time is needed under such high resolutions, which slowdowns analysis leading to lacking sufficient data points (Fig. 10.8).

10.2.6.5 Tandem Mass Spectrometry (MS/MS)

In conventional electrospray ionization (ESI) and nano-ESI, both basic peptides and proteins would be ionized with multiple charges higher than 3, which highly complicate assignment of their daughter ions. The fragmentation technologies of electron transfer dissociation (ETD) and electron capture dissociation (ECD) can reduce the charge of precursors and can be applied to both top-down and middle-down proteomic analyses.

Isomers with symmetric or asymmetric dimethylated arginines are indistinguishable in MS but in MS/MS. However, even by using FT-ICR MS and ETD and ECD MS/MS, it is challenging to solve variety of histone tail modifications completely. A higher performance of MS/MS in sensitivity and resolution are essential to determine modification sites. When one peptide containing both acetylation and tri-methylation, their sites can be determined only with a MS/MS resolution similar to MS. Since the sensitivity and resolution of MS/MS is in principle lower than those of MS, it is desirable for analysis of various histone-tail modifications to improve MS/MS in sensitivity, resolution and speed much further.

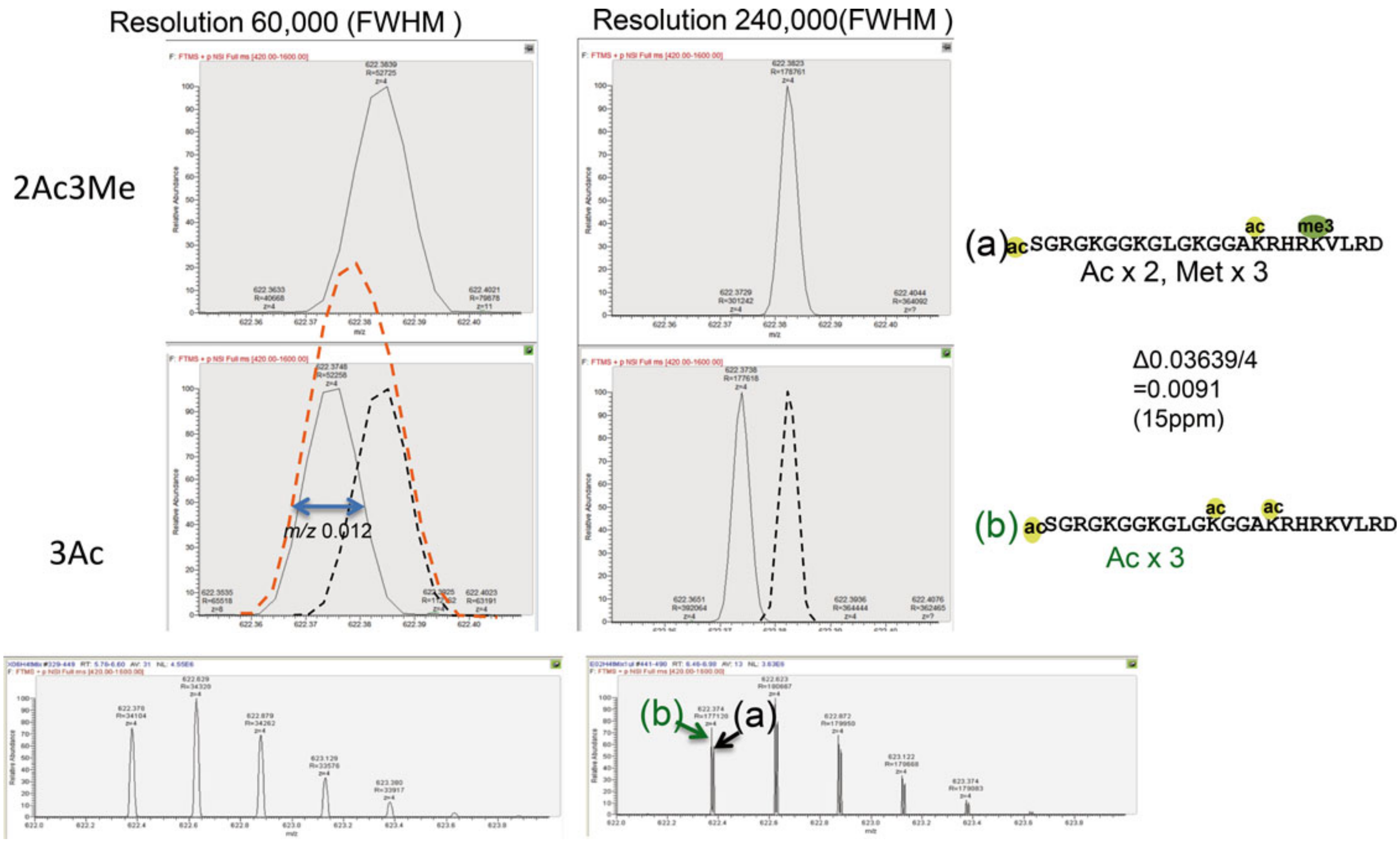


Fig. 10.8 Mass resolutions and modification. Difference of Ac and me3 are distinguishable at 240,000 resolution but not at 60,000

10.2.6.6 Database Search

A conventional proteomic database search takes into account merely multiple modifications on single peptide. In histone analysis, as in H4 acetylation that could occur on N-term, K5, 8, 12, 16, and 20, 6 acetylation sites can be simultaneously detected in a mixed sample even in case of mono-acetylation. Although these analyses can be performed by MS/MS, the site localization of modifications remains still difficult to be investigated because of sequence-specific fragmentations even with further improvement of MS instruments are achieved. In the previous work, localization had been determined by the manual assignment or using in-house software, and so methods are not generalized (Phanstiel et al. 2008; Young et al. 2009).

Searching the database for multiple modifications consumes much time. For example, in the case of Mascot database search (under the MS tolerance of 5 ppm, MS/MS tolerance of 0.8 Da, miss cleavage 0, fragmentation type ETD against Swissprot human), the searching time is almost the same between no modification and with protein N-term acetyl and Oxidation M. However, time would be taken more than 100 times when the epigenetic modifications of acetylation, mono, di, tri-methyl methylation, citrullination and phosphorylation are considered for searching.

It would be problematic that the risk of misidentification increases with increasing possible modifications, not that more searching time is taken. Some modifications have the same mass shift as amino acid substitutions such that methylation equals Ala/Gly and Val/Leu or Ile substitution, citrulination and amidation equal Asn/Asp substitution. This misidentification might be the issue for the analysis of histone H2 with many variants.

Above 15 kinds of modifications are set for the complete analysis of the H3 tail, but no searching software that accepts such big number of searching parameters is available. Ubiquitination is an important epigenetic mark and is detected as an adduct of GG at lysine residue in trypsin digest, but other enzymes than trypsin does not digest the long ubiquitin side chain, and so middle down analysis cannot be applicable on ubiquitination.

Among MS/MS spectra acquired, only half of them can be assigned to peptides with reliable scores. Unidentified spectra might contain undefined modifications. In fact, when unknown molecular shifts are taken into account, the number of identification would be improved considerably. Regarding to purified proteins, with considering unknown molecular shift and mis cleavage of protease, their sequence coverage sometimes increases to nearly 100 %. Therefore, unidentified modifications can be clarified when their unknown molecular shifts can be converted to chemical compositions.

10.2.6.7 Quantification

Quantification of epigenome is highly challenging compared to that of proteome. The stable isotope labeling using amino acids in cell culture (SILAC) or label free semi-quantification are utilized that are based on the measurement of ion-peaks separated. Both are applicable for histone modification but the separation of isobaric ions in LC are restricted. Then, quantification based on MS/MS, that is the selected-reaction monitoring (SRM)/multiple-reaction monitoring (MRM), are conducted (Zheng et al. 2012, 2013). We have demonstrated a successful determination of the acetylation sites by using both SRM and ion mobility approaches although some issues still remain for quantification.

10.2.6.8 Analysis of Histone H4 and H3 Tails

The histone H4 tail has been analyzed successfully. Coon et al. performed the comprehensive analysis of H4 tail in the ES differentiation, and found 73 patterns of modification (Phanstiel et al. 2008). Garcia et al. reported the H4-tail dynamics after adenovirus infection (ASMS 2013). We also utilized with the similar methods to analyze dynamics of a combinatorial H4 tail modification in cell cycle (unpublished). There, the dynamical combinations of modifications, including previously known H4K20 monomethylation at G2/M, has been elucidated, where AspN is commonly used since trypsin is not suitable for the analysis of combinatorial modifications (Fig. 10.9).

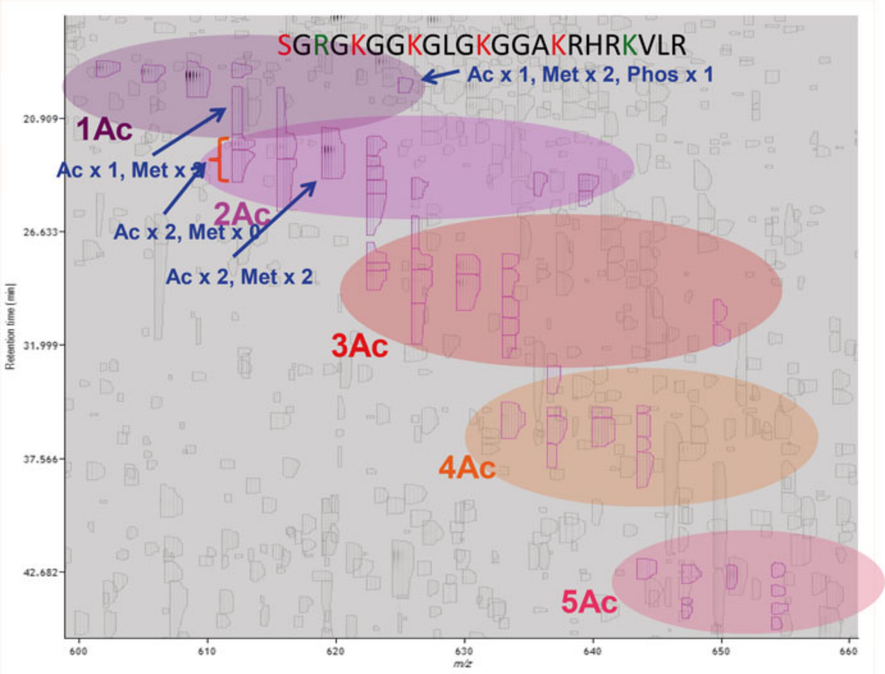


Fig. 10.9 Mass spectrometry analysis of H4 tail. 2D-image of the MS data separated and measured by nano RF-HPLC and Orbitrap. Progensis software (Nonlinear) was used. Peaks that contains different combinatorial modifications are separated

H3 analysis is more difficult than that of H4. As in the case of H4 analysis, lysine propionylation and trypsin digestion is not suitable. Therefore, top-down or middle-down using the enzymatic cleavage by GluC can be applied to analyze the whole H3 tail (Garcia et al. 2008; Kalli et al. 2013). In the top-down analysis, purification by modifications prior to LC/MS/MS experiments would be useful for successful analyses (Yates and Kelleher 2013). Middle-down approaches enable us to conduct a direct LC/MS/MS experiments. Such a comprehensive middle-down analysis would be soon available.

10.2.6.9 Perspective

In order to investigate disease-related marginal changes in histone profile in clinical samples, mass spectrometric performances on resolution, accuracy, and speed are expected to be improved furthermore. Since the ETD/ECD fragmentation needs more time than CID, the faster ETD/ECD is desirable. MS/MS in CID ought to be preferably performed under higher resolution to analyze details of fragment ions with multiple charges. Such a instrumentation of MS is not attainable at present but in near future their capability would be greatly improved. Concerning with hardware development, high quality data will be obtained by using highly advanced algorithms. The middle-down and top-down approaches are ineludible in profiling histone tails but a software able to handle their outputs has not been well developed yet, and at present both have several issues needed to be solved, such as determination of monoisotopic masses of precursor ions and deconvolution of fragment ions. Future advancement of those approaches would definitely bring a comprehensive analysis in proteomics as in the genomics to transomics.

10.2.7 Summary

Epigenetic changes are reversible. Understanding epigenome changes in diseases and restoring the changes would be highly expected to cure various types of diseases. The HMT and HDM, consisting respectively of 60 and 40 kinds, would be the therapeutic targets that are most likely to be effective to control accurate regulations. It should be noted that a strategic targeting a combination of reader proteins recognizing epigenetic marks and/or histone crosstalk factors might explore the drug discovery of next generation.

References

Ando T, Yoshida T, Enomoto S, Asada K, Tatematsu M, Ichinose M, Sugiyama T, Ushijima T. DNA methylation of microRNA genes in gastric mucosae of gastric cancer patients: its possible involvement in the formation of epigenetic field defect. Int J Cancer. 2009;124:2367–74.

Anway MD, Skinner MK. Epigenetic programming of the germ line: effects of endocrine disruptors on the development of transgenerational disease. Reprod Biomed Online. 2008;16:23–5.

Anway MD, Cupp AS, Uzumcu M, Skinner MK. Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science. 2005;308:1466–9.

Arnaudo AM, Garcia BA. Proteomic characterization of novel histone post-translational modifications. Epigenetics Chromatin. 2013;6:24.

Copeland RA, Solomon ME, Richon VM. Protein methyltransferases as a target class for drug discovery. Nat Rev Drug Discov. 2009;8:724–32.

Dai C, Gu W. p53 post-translational modification: deregulated in tumorigenesis. Trends Mol Med. 2010;16:528–36.

David Allis C, Jenuwein T, Reinberg D. Epigenetics. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2007.

Denell RE. Homoeosis in Drosophila. I. Complementation studies with revertants of Nasobemia. Genetics. 1973;75:279–97.

Di Leva G, Croce CM. miRNA profiling of cancer. Curr Opin Genet Dev. 2013;23:3–11.

Dvash T, Fan G. Epigenetic regulation of X-inactivation in human embryonic stem cells. Epigenetics. 2009;4:19–22.

Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358:1148–59.

Fischle W, Wang Y, Allis CD. Binary switches and modification cassettes in histone biology and beyond. Nature. 2003;425:475–9.

Fujita T, Fujii H. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. Biochem Biophys Res Commun. 2013;439:132–6.

Garcia BA, Thomas CE, Kelleher NL, Mizzen CA. Tissue-specific expression and post-translational modification of histone H3 variants. J Proteome Res. 2008;7:4225–36.

Graff J, Tsai LH. Histone acetylation: molecular mnemonics on the chromatin. Nat Rev Neurosci. 2013;14:97–111.

Hipolito CJ, Suga H. Ribosomal production and in vitro selection of natural product-like peptidomimetics: the FIT and RaPID systems. Curr Opin Chem Biol. 2012;16:196–203.

Jurinke C, Denissenko MF, Oeth P, Ehrich M, van den Boom D, Cantor CR. A single nucleotide polymorphism based approach for the identification and characterization of gene expression modulation using MassARRAY. Mutat Res. 2005;573:83–95.

Kalli A, Sweredoski MJ, Hess S. Data-dependent middle-down nano-liquid chromatography-electron capture dissociation-tandem mass spectrometry: an application for the analysis of unfractionated histones. Anal Chem. 2013;85:3501–7.

Kinoshita T, Ikeda Y, Ishikawa R. Genomic imprinting: a balance between antagonistic roles of parental chromosomes. Semin Cell Dev Biol. 2008;19:574–9.

Kipp DR, Quinn CM, Fortin PD. Enzyme-dependent lysine deprotonation in EZH2 catalysis. Biochemistry. 2013;52:6866–78.

Kon A, Shih LY, Minamino M, Sanada M, Shiraishi Y, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, Shimamura T, Tanaka H, Chiba K, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, Miyano S, Ogawa S. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. Nat Genet. 2013;45:1232–7.

Kooistra SM, Helin K. Molecular mechanisms and potential functions of histone demethylases. Nat Rev Mol Cell Biol. 2012;13:297–311.

Kubicek S, O'Sullivan RJ, August EM, Hickey ER, Zhang Q, Teodoro ML, Rea S, Mechtler K, Kowalski JA, Homon CA, Kelly TA, Jenuwein T. Reversal of H3K9me2 by a small-molecule inhibitor for the G9a histone methyltransferase. Mol Cell. 2007;25:473–81.

Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D. Histone methyltransferase activity associated with a human multiprotein complex containing the enhancer of Zeste protein. Genes Dev. 2002;16:2893–905.

Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell. 2013;152:1308–23.

Lorenz KZ. The evolution of behavior. Sci Am. 1958;199:67–74. passim.

Maekita T, Nakazawa K, Mihara M, Nakajima T, Yanaoka K, Iguchi M, Arii K, Kaneda A, Tsukamoto T, Tatematsu M, Tamura G, Saito D, Sugimura T, Ichinose M, Ushijima T. High levels of aberrant DNA methylation in Helicobacter pylori-infected gastric mucosae and its possible association with gastric cancer risk. Clin Cancer Res. 2006;12:989–95.

Marazzi I, Ho JS, Kim J, Manicassamy B, Dewell S, Albrecht RA, Seibert CW, Schaefer U, Jeffrey KL, Prinjha RK, Lee K, Garcia-Sastre A, Roeder RG, Tarakhovsky A. Suppression of the antiviral response by an influenza histone mimic. Nature. 2012;483:428–33.

McCabe MT, Ott HM, Ganji G, Korenchuk S, Thompson C, Van Aller GS, Liu Y, Graves AP, Della Pietra 3rd A, Diaz E, LaFrance LV, Mellinger M, Duquenne C, Tian X, Kruger RG, McHugh CF, Brandt M, Miller WH, Dhanak D, Verma SK, Tummino PJ, Creasy CL. EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. Nature. 2012;492:108–12.

Mohammed SI, Springfield S, Das R. Role of epigenetics in cancer health disparities. Methods Mol Biol. 2012;863:395–410.

Muller C, Leutz A. Chromatin remodeling in development and differentiation. Curr Opin Genet Dev. 2001;11:167–74.

Nair SS, Kumar R. Chromatin remodeling in cancer: a gateway to regulate gene transcription. Mol Oncol. 2012;6:611–19.

Ng SF, Lin RC, Laybutt DR, Barres R, Owens JA, Morris MJ. Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring. Nature. 2010;467:963–6.

Nishiyama A, Yamaguchi L, Sharif J, Johmura Y, Kawamura T, Nakanishi K, Shimamura S, Arita K, Kodama T, Ishikawa F, Koseki H, Nakanishi M. Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication. Nature. 2013;502:249–53.

Nozawa RS, Nagao K, Igami KT, Shibata S, Shirai N, Nozaki N, Sado T, Kimura H, Obuse C. Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. Nat Struct Mol Biol. 2013;20:566–73.

Peters AH, Kubicek S, Mechtler K, O'Sullivan RJ, Derijck AA, Perez-Burgos L, Kohlmaier A, Opravil S, Tachibana M, Shinkai Y, Martens JH, Jenuwein T. Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol Cell. 2003;12:1577–89.

Phanstiel D, Brumbaugh J, Berggren WT, Conard K, Feng X, Levenstein ME, McAlister GC, Thomson JA, Coon JJ. Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. Proc Natl Acad Sci U S A. 2008;105:4093–8.

Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007;447:425–32.

Richly H, Aloia L, Di Croce L. Roles of the polycomb group proteins in stem cells and cancer. Cell Death Dis. 2011;2:e204.

Rivera CM, Ren B. Mapping human epigenomes. Cell. 2013;155:39–55.

Rodriguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nat Med. 2011;17:330–9.

Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet. 2008;9:179–91.

Seong KH, Li D, Shimizu H, Nakamura R, Ishii S. Inheritance of stress-induced, ATF-2-dependent epigenetic change. Cell. 2011;145:1049–61.

Shin T, Kraemer D, Pryor J, Liu L, Rugila J, Howe L, Buck S, Murphy K, Lyons L, Westhusin M. A cat cloned by nuclear transplantation. Nature. 2002;415:859.

Singh TR, Gupta A, Suravajhala P. Challenges in the miRNA research. Int J Bioinform Res Appl. 2013;9:576–83.

Szyf M, Detich N. Regulation of the DNA methylation machinery and its role in cellular transformation. Prog Nucleic Acid Res Mol Biol. 2001;69:47–79.

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009;324:930–5.

Tan J, Yang X, Zhuang L, Jiang X, Chen W, Lee PL, Karuturi RK, Tan PB, Liu ET, Yu Q. Pharmacologic disruption of polycomb-repressive complex 2-mediated gene repression selectively induces apoptosis in cancer cells. Genes Dev. 2007;21:1050–63.

Tan YR, Peng D, Chen CM, Qin XQ. Nonstructural protein-1 of respiratory syncytial virus regulates HOX gene expression through interacting with histone. Mol Biol Rep. 2013;40:675–9.

Turek-Plewa J, Jagodzinski PP. The role of mammalian DNA methyltransferases in the regulation of gene expression. Cell Mol Biol Lett. 2005;10:631–47.

Tweedie-Cullen RY, Brunner AM, Grossmann J, Mohanna S, Sichau D, Nanni P, Panse C, Mansuy IM. Identification of combinatorial patterns of post-translational modifications on individual histones in the mouse brain. PLoS One. 2012;7:e36980.

Waddington C. The epigenotype. Endeavour. 1942;1:18–20.

Wagner JM, Hackanson B, Lubbert M, Jung M. Histone deacetylase (HDAC) inhibitors in recent clinical trials for cancer therapy. Clin Epigenetics. 2010;1:117–36.

Waldmann T, Schneider R. Targeting histone modifications–epigenetics in cancer. Curr Opin Cell Biol. 2013;25:184–9.

Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT, Wu X, Stack EC, Loda M, Liu T, Xu H, Cato L, Thornton JE, Gregory RI, Morrissey C, Vessella RL, Montironi R, Magi-Galluzzi C, Kantoff PW, Balk SP, Liu XS, Brown M. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. Science. 2012;338:1465–9.

Yates 3rd JR, Kelleher NL. Top down proteomics. Anal Chem. 2013;85:6151.

Yost JM, Korboukh I, Liu F, Gao C, Jin J. Targets in epigenetics: inhibiting the methyl writers of the histone code. Curr Chem Genomics. 2011;5:72–84.

Young NL, DiMaggio PA, Plazas-Mayorca MD, Baliban RC, Floudas CA, Garcia BA. High throughput characterization of combinatorial histone codes. Mol Cell Proteomics. 2009;8:2266–84.

Young NL, Dimaggio PA, Garcia BA. The significance, development and progress of high-throughput combinatorial histone code analysis. Cell Mol Life Sci. 2010;67:3983–4000.

Zempleni J, Chew YC, Hassan YI, Wijeratne SS. Epigenetic regulation of chromatin structure and gene function by biotin: are biotin requirements being met? Nutr Rev. 2008;66 Suppl 1:S46–8.

Zheng Y, Sweet SM, Popovic R, Martinez-Garcia E, Tipton JD, Thomas PM, Licht JD, Kelleher NL. Total kinetic analysis reveals how combinatorial methylation patterns are established on lysines 27 and 36 of histone H3. Proc Natl Acad Sci U S A. 2012;109:13549–54.

Zheng Y, Thomas PM, Kelleher NL. Measurement of acetylation turnover at distinct lysines in human histones identifies long-lived acetylation sites. Nat Commun. 2013;4:2203.

Chapter 11
Pharmacogenomics in Drug Development

Lena Gustavsson

Abstract Pharmacogenomics is the study on how variations in human genetics are influencing drug response. This includes variability in DNA and RNA like single nucleotide polymorphisms, gene duplications, epigenetics and gene expression. Extensive interindividual differences in drug response and toxicity are observed in most disease areas and pharmacogenomics is one of the factors leading to this variability. In drug development, understanding the pharmacogenetic impact on pharmacokinetics of new development compounds is important to properly design clinical studies and to optimize the treatment paradigm for specific patient populations. Drug metabolizing enzymes and transporters, that are key determinants of drug disposition, are commonly polymorphic. Dosing of drugs to individuals carrying a polymorphic allele encoding a reduced or increased metabolic/transport activity may lead to altered plasma and/or organ concentrations of the compound thereby leading to lack of efficacy or adverse drug reactions. This chapter is focused on pharmacogenetic factors that influence the pharmacokinetics of drugs with examples from therapies in the cancer, analgesia and cardiovascular areas.

Keywords Pharmacogenomics • DNA • Drug metabolism • Drug transporter • CYP450 • OATP1B1 • Therapy • Polymorphism

11.1 Introduction

Pharmacogenomics is the study on how variations in human genetics are influencing drug response. This includes variability in DNA and RNA like single nucleotide polymorphisms, gene duplications, epigenetics and gene expression. Extensive inter-individual differences in drug response and toxicity are observed in most disease areas and pharmacogenomics is one of the key determinants of this variability. Taking pharmacogenomic biomarkers into account to predict the pharmacological efficacy and risk of adverse events constitutes a great potential to design the optimal

L. Gustavsson (✉)
Department of Drug Metabolism, H. Lundbeck A/S, Valby, Denmark
e-mail: LEGU@lundbeck.com

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_11

treatment for each individual patient (Sim and Ingelman-Sundberg 2011). Technological advances such as development of efficient genome sequencing techniques have dramatically increased the possibilities to identify new genetic variants and to provide tools to screen for genetic biomarkers in large patient materials for the use of pharmacogenetic information in clinical practice. However, there is also a need to translate the immense genetic information generated to what is clinically relevant and has an impact on therapeutic response and adverse events.

Adverse events as a result of drug therapy is a major problem which causes substantial morbidity, mortality and healthcare costs (Davies et al. 2009). It has been reported that around 7 % of all hospital admissions and 30 % of admissions in the elderly population are caused by adverse drug events (Davies et al. 2009; Paul et al. 2008). Interindividual variability in pharmacokinetics is a major cause of adverse drug reactions (Sim and Ingelman-Sundberg 2011). Drug-drug interactions and pharmacogenomics are two important factors that may influence the plasma and/or organ concentrations of a drug and/or metabolites and consequently induce toxic responses. Pharmacogenomics provide a great potential in selecting the optimal drug for each individual patient to achieve pharmacological efficacy and to decrease the frequency of adverse drug reactions. This is of highest importance for drugs that have a narrow therapeutic window.

In drug development, prediction of potential variability in pharmacokinetics and therapeutic response is important. Understanding the pharmacogenetic impact on pharmacokinetics, pharmacodynamics and adverse drug reactions of new development compounds is key to properly design clinical studies and to optimize the treatment paradigm for a specific patient population. During the later stages of drug development pharmacogenomics information is being included into the product label. To guide the pharmaceutical industry, regulatory authorities like the US Food and Drug Federation (FDA) and the European Medicines Agency (EMA) have during the last decade issued guidelines in this area. Pharmacogenomic biomarkers may be divided into pharmacodynamic, that are based on differences in the molecular target and disease characteristics, or pharmacokinetic, that are based on differences in drug metabolizing enzymes and transporters that determine drug disposition (Sim and Ingelman-Sundberg 2011). Drug disposition determines the concentration of drug at the target site as well as other organs and is therefore tightly linked to the pharmacological efficacy and toxicity. This chapter is focused on pharmacogenetic factors that influence the pharmacokinetics of drugs with examples from therapies in the cancer, analgesia and cardiovascular areas.

11.2 Determinants of Drug Disposition

Drug disposition is the fate of the molecule after it enters the body. After oral dosing, this includes the absorption of the drug across the intestinal epithelium into the systemic circulation, distribution of the drug to and between tissues, the metabolism of the drug and the excretion of drug most commonly into urine and faeces. The

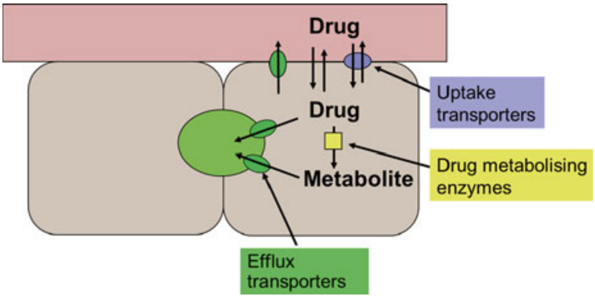


Fig. 11.1 Different processes involved in drug disposition as illustrated by drug transport and metabolism in hepatocytes. The drug is taken up to the hepatocytes by passive diffusion or transporter proteins. The drug may be metabolized by phase I and/or phase II enzymes. The parent drug and its metabolites may be translocated back to blood by active and passive mechanisms as well as being secreted out in the bile. A majority of drug metabolizing enzymes and transporters are subject to genetic polymorphism which causes interindividual differences in drug disposition

pharmacokinetics of a drug is partly determined by its physicochemical characteristics, e.g. lipophilicity, that are important for the passive diffusion across biological membranes and partly dependent on the compound's affinity for drug metabolizing enzymes and transporters (Fig. 11.1). Physiological characteristics like the blood flow and perfusion rates to different organs influence how the drug is being distributed. Drug metabolizing enzymes and transporters are commonly polymorphic and their activity is consequently influenced by the genotype. Absorption across the intestinal epithelium may occur as passive diffusion but there are also transporter proteins in the epithelial cells that may hinder or facilitate the absorption process. In addition to passive processes like diffusion and affinity for e.g. plasma proteins, drug distribution in the body is governed by drug transporters. Excretion of drugs into urine may occur passively through glomerular filtration but is also regulated by drug transporters mediating the secretion of drugs across the proximal tubule cells in the kidney. Biliary excretion of drugs and their metabolites are also driven by transporter proteins in the hepatocytes.

11.3 Drug Metabolizing Phase I Enzymes and Polymorphism

The majority of drug metabolism occurs in the liver and commonly converts the drug to a more hydrophilic species that is readily excreted. Drug metabolism is mediated by phase I and phase II enzymes. Phase I enzymes are catalyzing oxidation, reduction or hydrolysis of xenobiotics. The Cytochrome P450 (CYP) family is the most important family of phase I enzymes. Phase II enzymes are carrying out conjugation reactions, adding a hydrophilic moiety like a glucuronic acid or sulphate to the drug or its metabolites.

Of the more than 50 P450 isoforms expressed in the human body, around 9 isoforms belonging to the CYP1A, CYP2A, CYP2B, CYP2C, CYP2D and CYP3A families are the most important in the metabolism of drugs (Wienkers and Heath 2005). Several of these P450 enzymes are polymorphic. Genetic variants of P450s include single nucleotide polymorphisms (SNPs) that cause a change in the amino acid sequence and may cause altered substrate specificity as well as a changed activity (Eichelbaum et al. 2006; Ingelman-Sundberg et al. 2007). Some of the SNPs result in null alleles that are not translated into the active protein due to defect splicing. SNPs in the promoter region have also been found and these mutations may lead to an increased enzyme activity. An increased P450 activity is also commonly occurring as a result of increased number of gene copies (Johansson and Ingelman-Sundberg 2008). There are large interethnic variations in the polymorphic allele frequencies as illustrated in Table 11.1. The different CYP alleles are summarized at the Human CYP Allele Nomenclature Committee homepage (www.cypalleles.ki.se). Variability in the activity of CYP enzymes may also be caused by epigenetic factors and by induction but these mechanisms will not be covered in this chapter.

11.3.1 CYP2D6

CYP2D6 is one of the major drug metabolizing enzymes being important in the metabolism of around 25 % of drugs on the market (Eichelbaum et al. 2006). CYP2D6 catalyzes the metabolism of several antidepressants, neuroleptics, analgesics and anticancer drugs. CYP2D6 is highly polymorphic and allelic variants include several single nucleotide polymorphisms (Ingelman-Sundberg et al. 2007) (Table 11.1). In addition to null alleles and allelic variants resulting in reduced activity, CYP2D6 is also subject to copy number variations resulting in an increased enzyme activity (Ingelman-Sundberg et al. 2007). CYP2D6 phenotypes are generally divided into four groups; (1) poor metabolizers that are homozygous for null alleles, (2) intermediate metabolizers that either are heterozygous for null alleles or homozygous for alleles encoding defective protein, (3) extensive metabolizers that are referred to as the normal CYP2D6 activity with two functional alleles and (4) ultra-rapid metabolizers that display an increased CYP2D6 activity due to gene duplication/multiplication (Ingelman-Sundberg and Sim 2010). Poor metabolizers constitute a significant proportion of the population with 5–10 % of Caucasians (Table 11.1) and this is important to take into account when dosing drugs that are dependent on CYP2D6 metabolism for their elimination or activation. Equally well may ultra-rapid metabolizers be subject to lack of efficacy or adverse events due to a rapid inactivation of drugs and/or rapid metabolism to a toxic metabolite. CYP2D6 is the pharmacogenomic biomarker with the highest number of assignments in drug labels (Frueh et al. 2008). Several antipsychotics and antidepressants are substrates of CYP2D6 and the plasma concentration of these may vary 5–20-fold between individuals. In drug development, it is of great importance to understand the contribution of CYP2D6 to the metabolism of new compounds and pharmacogenetic

Table 11.1 Selected genetic polymorphisms of P450 enzymes and their frequency in different populations

Gene	rs number	Nucleotide change	Effect	Function	Clinical impact – examples	Allele frequency (%)		
						Caucasians	Africans	Asians
CYP2A6								
CYP2A6*2	rs1801272	1799T>A	L160H	No activity	↓ nicotine metabolism, ↑ nicotine dependence	4–10	0–1	0–3
CYP2C9								
CYP2C9*2	rs1799853	3608C>T	R144C	↓ activity	Warfarin: ↓CL and ↑ADR (bleeding)	10–17	0–2	0–2
					NSAIDs: ↓CL & ↑ADR			
CYP2C9*3	rs1057910	42614A>C	I359L	↓ activity	Similar to CYP2C9*2	6	~0	2–6
CYP2C19								
CYP2C19*2	rs4244285	19154G>A,	Splicing defect	Null allele	Clopidogrel: ↓ effect and ↑ cardiovascular events	6–15	10–17	22–32
					Antidepressants: ↓ CL and ↑ ADR			
CYP2C19*3	rs4986893	17948G>A	W212X	Null allele	Same as CYP2C19*2	0–1	0–1	3–7
CYP2C19*17	rs12248560	806C>T	Promoter	↑ activity and expression	Clopidogrel:↑bleeding risk	21–25	15–27	0–2
CYP2D6								
CYP2D6*2xn	–	Gene duplication/ multiplication	Copy number variations	↑ activity	Antidepressants: ↑ CL, ↓ effect	1–9	1–16	~1
					Codeine: ADR			
CYP2D6*4	rs3892097	1846G>A	Splicing defect	Null allele	Antidepressants: ↓ CL and ↑ADR tamoxifen: ↓ effect	11–29	1–4	0–1
					Codeine: ↓ effect			
CYP2D6*5	–	Gene deletion		Null allele	Similar to CYP2D6*4	1–7	3–6	1–6
CYP2D6*10	rs1065852	100C>T	P34S	Unstable enzyme	Similar to CYP2D6*4	1–6	4–9	38–70
CYP2D6*17	rs28371706 rs16947	1023C>T, 2850C>T	T107I, R296C	Altered substrate affinity	Similar to CYP2D6*4 but less pronounced	0–1	9–34	0

The information is collected from Eichelbaum et al. (2006), Ingelman-Sundberg and Sim (2010), Zanger and Schwab (2013) and www.cypalleles.ki.se (*CL* clearance, *ADR* adverse drug reaction)

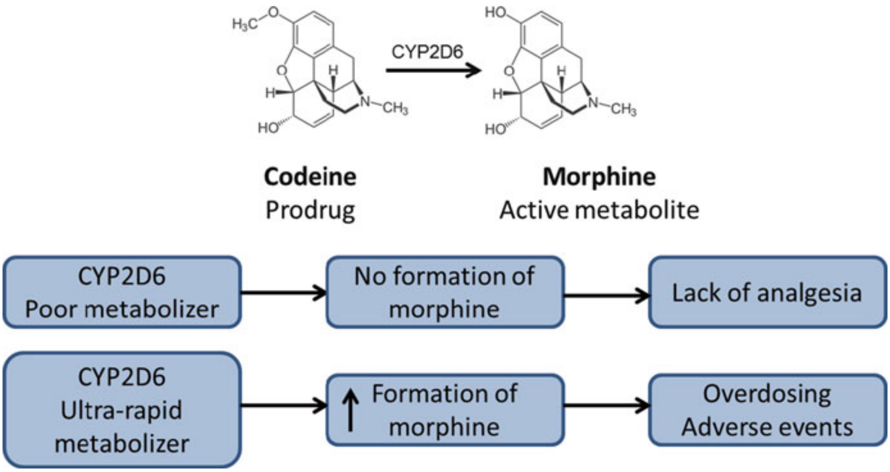


Fig. 11.2 Metabolism of codeine to its active metabolite morphine catalyzed by CYP2D6. Polymorphic variants lead to either a poorly metabolizing phenotype or an ultra-rapid phenotype that results in lack of pharmacological efficacy or adverse drug events, respectively

studies to investigate the impact are recommended by regulatory authorities if CYP2D6 is predicted to contribute extensively to the elimination of a drug (Maliepaard et al. 2013).

A well-established example of pharmacogenetic impact by CYP2D6 polymorphism is the variability in the analgesic effect of codeine. Codeine is a prodrug metabolized by CYP2D6 to the pharmacologically active metabolite morphine which has a 200 times higher affinity to the μ-opioid receptor (Ingelman-Sundberg et al. 2007) (Fig. 11.2). Poor metabolizers of CYP2D6 are unable to form morphine from codeine and consequently experience a decrease in the analgesic effect. On the other hand, ultrarapid metabolizers showed 50 % higher plasma concentrations of morphine than extensive metabolizers after codeine administration (Kirchheiner et al. 2007). The increased rate of morphine formation in ultrarapid metabolizers may potentially cause adverse effects like extreme sleepiness, confusion and shallow breathing (Crews et al. 2014). Particular attention has been drawn to consider pharmacogenomics of CYP2D6 in the dosing of codeine to breastfeeding women (Frueh et al. 2008; Crews et al. 2014; Daly 2010).

Tamoxifen is a selective estrogen receptor (ER) antagonist widely used for the treatment of ER-positive breast cancer. However, in as much as one third of the patients the disease recurs. Tamoxifen is a prodrug and is activated by P450 mediated metabolism including CYP2D6 and CYP3A4 as key enzymes (Fig. 11.3) (Desta et al. 2004; Singh et al. 2011). Importantly, CYP2D6 is the rate-limiting step in the metabolism of tamoxifen to the active metabolite endoxifen that has a 100-fold higher affinity for the ER than the parent drug (Stearns et al. 2003). CYP2D6 is also involved in the formation of the pharmacologically active metabolite 4-hydroxytamoxifen which has a potency towards ER similar to endoxifen. Given that the activation of tamoxifen is dependent on a highly polymorphic enzyme, CYP2D6, several studies have investigated whether the

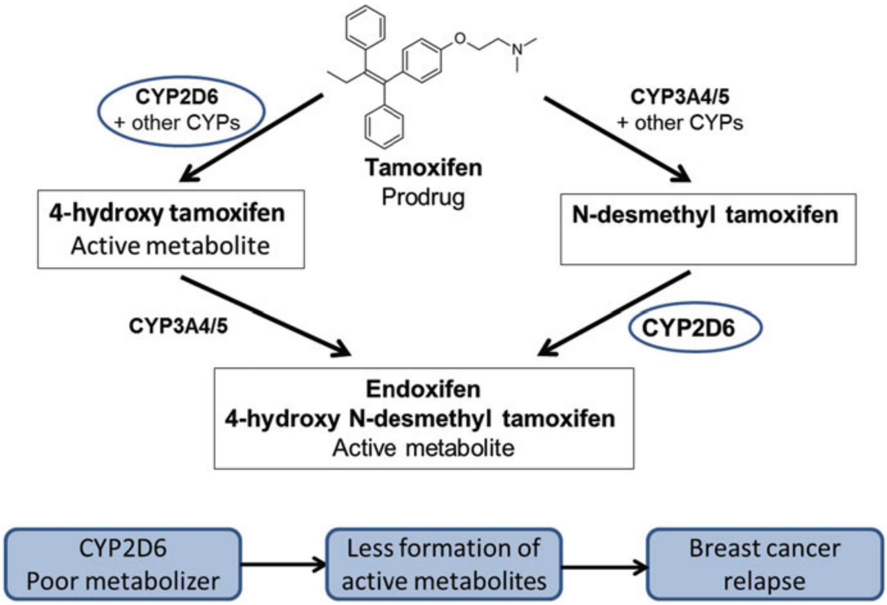


Fig. 11.3 Metabolism of tamoxifen to its active metabolites endoxifen and 4-hydroxy tamoxifen. The metabolism is dependent on CYP2D6 and poor metabolizers show lower plasma concentrations that may lead to increased risk for relapse

interindividual differences in pharmacokinetics and drug response could be explained by variability in CYP2D6 expression. Pharmacokinetic studies on tamoxifen and metabolite formation demonstrated that plasma concentrations of endoxifen were only around 25 % in CYP2D6 poor metabolizers compared to CYP2D6 extensive metabolizers (Borges et al. 2006). In addition to pharmacokinetic studies, there is also compelling clinical evidence that CYP2D6 poor metabolizers receiving tamoxifen treatment of breast cancer have an increase in relapse rate and lower survival than extensive metabolizers (Kiyotani et al. 2008, 2010, 2013).

11.3.2 CYP2C19

CYP2C19 is responsible for metabolism of 5–10 % of clinically used drugs including several antidepressants, proton pump inhibitors as well as the anticoagulant clopidogrel (Zanger and Schwab 2013). Similar to CYP2D6, CYP2C19 is highly polymorphic and two null alleles, CYP2C19*2 and CYP2C19*3, constitute the major allelic variants. Poor metabolizers are relatively frequent in some populations (Table 11.1) and results in a large variation in CYPC19 mediated clearance e.g. in the metabolism of clopidogrel.

Clopidogrel is a platelet P2Y$_{12}$ adenosine phosphate (ADP) receptor inhibitor used to decrease the risk of platelet aggregation in patients with coronary artery disease and other vascular disorders. Clopidogrel is a prodrug and formation of the pharmacologically active metabolite occurs in two steps both of them involving CYP2C19 (Kazui et al. 2010). Several studies and meta-analysis of data have demonstrated an association between CYP2C19 poor metabolizer phenotype and lower plasma concentrations of the active metabolite, decrease in platelet inhibition as well as higher rates of adverse events (Mega et al. 2011; Simon et al. 2011). Moreover, the polymorphic variant CYP2C9*17, known to encode for a CYP2C9 protein with higher than normal activity, has been associated with adverse drug reactions in the form of bleedings (Ingelman-Sundberg et al. 2007). Consequently, FDA has issued a boxed warning in the drug label to avoid clopidogrel dosing to CYP2C19 poor metabolizers (see link FDA homepage http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm).

11.3.3 CYP2C9

CYP2C9 is the one of the major CYP450s responsible for metabolism of 10–20 % of drugs on the market (Wienkers and Heath 2005). CYP2C9 metabolizes weakly acidic compounds including the anti-coagulant warfarin, non-steroidal anti-inflammatory drugs (NSAIDs), anticonvulsants like phenytoin and valproic acid, as well as the anti-diabetics glibenclamide and tolbutamide (Zanger and Schwab 2013). There are more than 50 polymorphic variants of CYP2C9 of which CYP2C9*2 and CYP2C9*3, that results in decreased activity, have been demonstrated to influence clinical pharmacokinetics of several substrates (Zanger and Schwab 2013). Several of the substrates of CYP2C9, e.g. warfarin, have a narrow therapeutic index which increases the risk for adverse side effects and sensitivity to variability in pharmacokinetics. In carriers of the heterozygous CYP2C9*3, clearance for several drugs was shown to be reduced to 40–75 % of the wild-type CYP2C9 carriers. A reduction to less than 25 % of oral clearance was demonstrated for homozygous CYP2C9*3 carriers (Kirchheiner and Brockmoller 2005).

Warfarin is one of the most widely prescribed oral anticoagulants used in the treatment and prevention of thrombolytic diseases. Warfarin is a coumarine derivative that specifically inhibits the vitamin K epoxide reductase (VKOR) encoded by the VKOR complex subunit 1 (VKORC1) gene (Yin and Miyata 2007). Thereby warfarin prevents the maturation of the vitamin K-dependent clotting factors. There are large inter-individual variations in the pharmacological response to warfarin and moreover, the drug has a narrow therapeutic index. Too low plasma concentrations of warfarin will fail to decrease the risk of thromboembolism whether a too high plasma concentration will increase the risk of bleeding. As both lack of efficacy and the adverse event in terms of bleeding may cause life-threatening events, the dosing has to be carefully controlled. Thus, during warfarin therapy the anticoagulative effect is regularly monitored by measuring the prothrombin time in each individual patient.

The dosing of warfarin is affected to approximately 40 % by genetic factors (Jonas and McLeod 2009). Two major factors influencing the effect of warfarin are the pharmacogenetic variants of the target VKORC1 and the drug metabolizing enzyme CYP2C9 (Yin and Miyata 2007; Jonas and McLeod 2009; Yang et al. 2013). Warfarin consists of two stereoisomers R-warfarin and S-warfarin. S-warfarin is fivefold more potent inhibitor of VKOR than the R-warfarin isomer and S-warfarin accounts for 60–70 % of the pharmacological response (Yin and Miyata 2007). S-warfarin is mainly metabolized by CYP2C9 whereas R-warfarin is metabolized by CYP1A1, CYP1A2 and CYP3A4. Patients carrying the CYP2C9*2 or CYP2C9*3 allele metabolize S-warfarin more slowly compared to the population carrying the wild-type gene (Yin and Miyata 2007). Using a normal dose to this group of patients is accompanied with a higher risk of bleeding and a dose reduction is therefore required. As an example, Sanderson et al (Sanderson et al. 2005) showed that a 30 % dose reduction in patients with the CYP2C9*3 allele was required to obtain efficacy without serious adverse events. There are indications that the CYP2C9 polymorphism has a larger impact on the initial dose finding period whereas it has less impact during long-term treatment. In particular during the period after start of dosing, patients with less active CYP2C9 variants require a longer time to achieve a stable dose and have a higher risk of serious or life-threatening bleeding events (Yin and Miyata 2007).

11.4 Drug Metabolizing Phase II Enzymes and Polymorphism

Several of the phase II enzymes are also polymorphic but the area of pharmacogenomics is still not as advanced for these enzymes as for the P450 families. Recently, Stingl and co-authors reviewed the pharmacogenomics of UDP-glucuronosyltransferases (UGT) and concluded that there are a large number of polymorphisms with influence on pharmacokinetics that potentially may be used for design of personalized treatment in the future (Stingl et al. 2014). The most well established pharmacogenomic biomarker in this area is the allelic variant UGT1A1*28 that encodes an enzyme with lower activity than the wild-type enzyme (Sim and Ingelman-Sundberg 2011; Stingl et al. 2014). The topoisomerase 1 inhibitor irinotecan, used in cancer therapy, is metabolized by carboxylesterase to its active metabolite SN-38. Inactivation and excretion of SN-38 is dependent on UGT1A1. Accumulation of SN-38 leads to serious adverse events in the form of neutropenia. Individuals who are homozygous for the UGT1A1*28 variant show a decreased clearance of SN-38 and therefore, carry a higher risk of neutropenia (Schulz et al. 2009). As a consequence, FDA recommends adjustment of the irinotecan dose to patients with UGT1A1*28.

11.5 Drug Transporters and Polymorphism

During the last ten year it has become evident that drug transporters play an important role in drug disposition affecting the absorption, distribution and excretion of drugs (Giacomini et al. 2010). Transporter proteins of the solute carrier family (SLC and SLCO) facilitate translocation of drugs across the plasma membrane most commonly in the inward direction. Transporters of the ATP-binding cassette family pump drugs and metabolites out of cells in a process driven by ATP. In 2010, the International Transporter Consortium (ITC), with participants from academia, regulatory authorities and pharmaceutical industry, published a comprehensive report on the clinical evidence of drug transporter function (Giacomini et al. 2010). Based on this review, regulatory authorities updated their guidelines to include those transporters that we know today are important in drug disposition. Recently, ITC published a commentary on drug transporters and pharmacogenomics (Giacomini et al. 2013). Similar to drug metabolizing enzymes, several drug transporters are polymorphic resulting in an interindividual variability in transport rate. Except for a few cases, there is however still limited evidence of their clinical relevance. Drug transporters may not only affect the plasma concentrations of drugs but also, and sometimes to a higher extent, the concentration of drug in specific organs. Thus a polymorphic variant that changes the activity of a transporter protein may result in drug and metabolite accumulation which potentially may lead to unpredicted toxicity. Drug concentrations in organs are difficult to measure in the clinic and thus mechanisms of toxicity due to transporter polymorphism may require advanced techniques such as non-invasive imaging.

The most well described polymorphic drug transporter is the organic anion transporting polypeptide OATP1B1 (SLCO1B1), see below. Also polymorphism of the breast cancer resistance protein (BCRP/ABCG2) has been demonstrated to impact the pharmacokinetics of its substrates. The breast cancer resistance protein (BCRP/ABCG2) belongs to the ATP-binding cassette family of transporters that are pumping compounds out of cells by an ATP dependent mechanism. BCRP has broad substrate specificity and transports a variety of drugs including chemotherapeutics, statins and antibiotics (Giacomini et al. 2013; Polgar et al. 2008). BCRP is expressed on the apical membrane of epithelial cells of the intestine where it may hinder drugs to get absorbed as well as in the liver and kidney where this transporter contributes to the excretion of drugs and their metabolites into bile and urine (Giacomini et al. 2013). BCRP is also expressed on the endothelial cells of the blood brain barrier where it, together with MDR1, restricts its substrates to enter the brain (Giacomini et al. 2013). In terms of polymorphism the allelic variant c.421C>A has been associated with increased plasma exposure of several of its substrates (Giacomini et al. 2013). As an example, the AUC of rosuvastatin was increased 2.4-fold in individuals homozygous for the ABCG2 c.421AA compared to the c.421CC genotype (Keskitalo et al. 2009).

11.5.1 OATP1B1

The organic anion transporting polypeptide 1 (OATP1B1/SLCO1B1) is highly expressed on the sinusoidal membrane of human hepatocytes where it facilitates the uptake of organic anions into the cells (Fig. 11.1). Several endogenous compounds such as bile salts, bilirubin glucuronides and steroid hormone metabolites as well as frequently used drugs like statins, HIV protease inhibitors, anti-diabetics and anticancer agents are substrates of OATP1B1 (Giacomini et al. 2010; Niemi et al. 2011). By taking up drugs into the hepatocytes, OATP1B1 is the first step in hepatic elimination of its substrates. For some drugs, in particular for compounds with a low passive permeability the transporter mediated uptake may be the rate limiting step in hepatobiliary excretion (Shitara et al. 2006).

OATP1B1 is a polymorphic gene and several of its allelic variants are associated with a modified transport activity (Tirona et al. 2001). There is compelling clinical evidence that polymorphism is an important factor in the interindividual variability of pharmacokinetics of OATP1B1 substrates (Giacomini et al. 2010). Interestingly, expression of OATP1B3 (SLCO1B3) which is structurally similar, has a large overlap in terms of substrate specificity with OATP1B1 and is expressed on the sinusoidal membrane of human hepatocytes, does not seem to be highly influenced by polymorphism (Nies et al. 2013). In particular, the OATP1B1 c.521T>C variant that is expressed in a substantial proportion of the population (Table 11.2), has been demonstrated to have a large impact on drug disposition. The c.521T>C variant encodes for a transporter protein with reduced activity towards several OATP1B1 substrates (Niemi et al. 2011). Consequently, due to a decreased uptake into the liver, the plasma concentration of OATP1B1 substrates are increased. In particular, individuals that are homozygous, carrying the c.521CC genotype are affected.

Although impact of pharmacogenomics of OATP1B1 on pharmacokinetics has been demonstrated for several drugs, most extensive studies have been reported on statins. Statins are inhibitors of 3-hydroxy-3-methyl-glutaryl coenzyme A (HMG-CoA) reductase that decrease the low-density lipoprotein cholesterol and are widely used to reduce the risk of cardiovascular disease. Statins are substrates of OATP1B1 and OATP1B3 with varying degree of substrate specificity (Sharma et al. 2012). They display a varying degree of physicochemical properties (Sharma et al. 2012) which influences their dependency on OATPs to enter their target cell which is the hepatocyte. As an example, rosuvastatin which has a log D at pH 7.4 of −0.3 has a poor passive permeability and consequently is highly dependent on a transporter to enter the cells whereas fluvastatin which has a log D of 1.4 has a higher passive permeability and the uptake transporter does therefore not have as big impact (Sharma et al. 2012).

Polymorphism in the OATP1B1 gene plays a key role in the inter-individual variability of pharmacokinetics of statins. The c.521T>C variant that occurs in several haplotypes (Table 11.2) and is associated with a reduced transport activity has

Table 11.2 Selected single-nucleotide polymorphisms and haplotypes of the OATP1B1 gene that are associated with a change in transport activity

				Allele frequency (%)			
	rs number	Nucleotide	Amino acid	Caucasians	Africans	East Asians	Transport activity
SNPs	rs2306283	c388A>G	p.N130D	40–41	81–83	66–80	Increase
	rs11045819	c.463C>A	p.P155T	14–19	2.2–6.8	0	Increase
	rs4149056	c.521T>C	p.V174A	15–22	0.7–12	12–14	Decrease
Haplotypes							
SLCO1B1*1b		c388A>G	p.N130D	11	73	62	Increase
SLCO1B1*5		c.521T>C	p.V174A	3	0	0	Decrease
SLCO1B1*15		c388A>G and c.521T>C	p.N130D and p.V174A	14	1.9	14	Decrease

Data are from Giacomini et al. (2013)

consistently been demonstrated to impact statin pharmacokinetics (Niemi et al. 2011). In agreement with a decrease in hepatic uptake, the nonrenal clearance of pravastatin was reduced in a population with the *15 haplotype (Nishizato et al. 2003) and in an additional study the area under the plasma concentration curve was increased in individuals carrying the c.521T>C genotype (Niemi et al. 2004). Different statins are affected by OATP1B1 polymorphism to different degrees. The largest effect of the c.521T>C genotype has been observed for simvastatin acid with a 3.2-fold increase in AUC in a homozygous population (Niemi et al. 2011). The plasma concentrations of atorvastatin, pravastatin and rosuvastatin were also significantly increased whereas fluvastatin AUC was not different in the c.521CC group as compared to individuals with the normal OATP1B1 genotype (Niemi et al. 2011). The difference may at least in part be explained by the dependence of each respective statin on OATP1B1 to be taken up into the liver. A higher degree of passive permeability and other transporters e.g. OATP1B3 may decrease the impact of OATP1B1 polymorphism.

The clinical implications of an increased statin concentration have been observed in terms of increased occurrence of adverse events. At high concentrations statins may cause myopathy which in rare cases accelerates to rhabdomyolysis (Ghatak et al. 2010). In a genome wide association study on a patient population with myopathy and simvastatin treatment, and a control population, the c.521CC genotype was shown to be strongly associated with myopathy (Link et al. 2008). This finding has been reproduced and milder adverse reactions have been found to be associated with the c.521C allele also for other statins (Voora et al. 2009). Consequently, the Clinical Pharmacogenomics Implementation Consortium (CPIC) has issued recommendations for genotyping and dosing adjustments for simvastatin (Wilke et al. 2012).

11.6 Pharmacogenomics in Drug Development

With the growing evidence of pharmacogenomics impact on pharmacokinetics, pharmacodynamics and adverse drug reactions, FDA did in 2005 issue a guidance in which information on pharmacogenomics is required for submission to regulatory authorities. Since then several additional documents and guidelines from FDA, EMA and Pharmaceuticals and Medical Devices Agency (PMDA) in Japan has been published in this area (for comprehensive list of documents, see (Maliepaard et al. 2013).

Before first dosing to man, major pathways of drug metabolism and transport are investigated to assess the risk of pharmacogenomic and drug-drug interaction factors in pharmacokinetics and pharmacodynamics (Fig. 11.4). If the drug is found to be metabolized or transported to a large extent by a specific enzyme/transporter protein, clinical studies have to be designed to investigate the impact on variability in pharmacokinetics. This will typically include investigation of pharmacokinetics

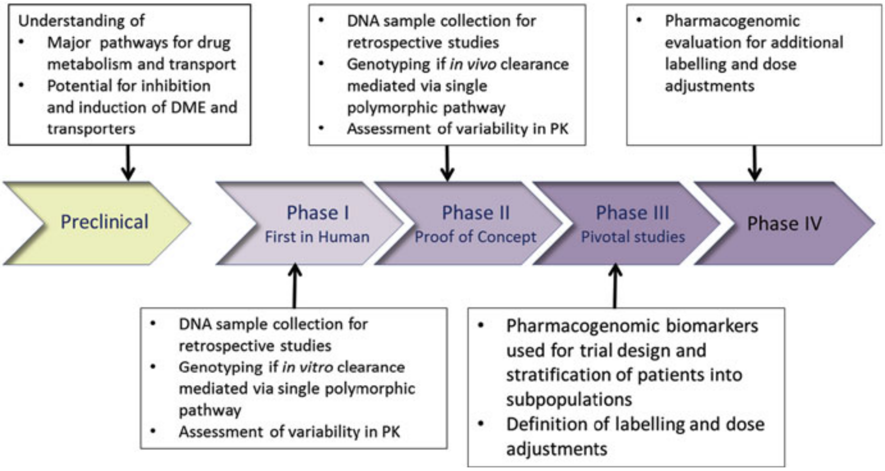

Understanding of
• Major pathways for drug metabolism and transport
• Potential for inhibition and induction of DME and transporters
• DNA sample collection for retrospective studies
• Genotyping if *in vivo* clearance mediated via single polymorphic pathway
• Assessment of variability in PK
• Pharmacogenomic evaluation for additional labelling and dose adjustments
Preclinical
Phase I
First in Human
Phase II
Proof of Concept
Phase III
Pivotal studies
Phase IV
• DNA sample collection for retrospective studies
• Genotyping if *in vitro* clearance mediated via single polymorphic pathway
• Assessment of variability in PK
• Pharmacogenomic biomarkers used for trial design and stratification of patients into subpopulations
• Definition of labelling and dose adjustments

Fig. 11.4 Pharmacogenomics for pharmacokinetic considerations during the different phases of drug development. *DME* drug metabolizing enzyme

of the parent drug and potentially the formation of specific metabolites in subpopulations with the allelic variant of interest compared to the normal genotype. In terms of pharmacogenomic impact, EMA states that genotyping is required during first time in man and further phase I studies if one single metabolic pathway is predicted from *in vitro* studies to be responsible for >50 % of the clearance of a compound (Maliepaard et al. 2013). It should be noted that FDA and PMDA do not publish any cut-off values and lower proportion of clearance than 50 % may thus require genotyping if appropriate. The pharmacogenetic data from early clinical studies may be used to lower the doses to be given to individuals that are poor metabolizers of the specific drug in later trials. Moreover, if *in vivo* clinical data indicate that >25 % of the drug is cleared by a single polymorphic pathway, genotyping is required during phase II studies. Similar requirements are valid for pharmacologically active metabolites. In general, collection of DNA samples is highly recommended from all clinical studies in order to enable retrospective evaluation of variabilities in pharmacokinetics and/or pharmacodynamics observed e.g. in phase I studies. Including pharmacogenomics data collection during early clinical studies may be important for formation of hypothesis for prospective studies to evaluate findings on variability in pharmacokinetics, pharmacological response and adverse drug reactions. Genotyping may also be highly valuable in forming the strategy and clinical study design to be used in phase III trials and may aid in stratification of patients into subpopulations with a high likelihood of being responders to a specific drug treatment and to avoid inclusion of subjects who may be more prone to toxic effects. Thus, this may enable selection of specific patient populations to focus large clinical trials on relevant patients.

11.7 Concluding Remarks and Future Perspectives

There is a great potential in pharmacogenomics to optimize drug therapy for the benefit of the patient at an improved cost – benefit balance. Genetic differences may impact on several aspects of disease and drug effect. This includes genes that governs the PK of a drug, genes encoding drug target and related pathways, genes predisposing to toxicity and genes influencing the sensitivity to disease.

Incorporating genotyping into drug therapy is a promising area in the development of personalized healthcare and may enhance the number of treatment responders while decreasing adverse events.

So far, pharmacogenomics has mainly had impact retrospectively when a drug has already been launched on the market. However, with the current increased awareness of the importance of genetic variability and association to drug response together with new high-throughput techniques for genotyping, there are enhanced prerequisites to use pharmacogenomics information prospectively during drug development. Thus, there is a high potential for new drugs to be developed for more highly defined patient populations and for which the dosing connected to pharmacological efficacy and avoidance of adverse side effects may be better controlled. However, there is also a need for large and well-designed prospective studies to validate new pharmacogenetic markers and to translate new findings into clinical practice. In current clinical practice, pharmacogenomics is only being used in a few cases. Product labelling is usually more informative than decision-making. Processes for implementation of a more efficient use of pharmacogenomics are required once the pharmacogenomics markers are well validated.

References

Borges S, Desta Z, Li L, et al. Quantitative effect of CYP2D6 genotype and inhibitors on tamoxifen metabolism: implication for optimization of breast cancer treatment. Clin Pharmacol Ther. 2006;80:61–74.

Crews KR, Gaedigk A, Dunnenberger HM, et al. Clinical pharmacogenetics implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. Clin Pharmacol Ther. 2014;95(4):376–82.

Daly AK. Pharmacogenetics and human genetic polymorphisms. Biochem J. 2010;429:435–49.

Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M. Adverse drug reactions in hospital in-patients: a prospective analysis of patient-episodes. PLoS One. 2009;4:e4439.

Desta Z, Ward BA, Soukhova NV, Flockhart DA. Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. J Pharmacol Exp Ther. 2004;310:1062–75.

Eichelbaum M, Ingelman-Sundberg M, Evans WE. Pharmacogenomics and individualized drug therapy. Annu Rev Med. 2006;57:119–37.

Frueh FW, Amur S, Mummaneni P, et al. Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. Pharmacotherapy. 2008;28:992–8.

Ghatak A, Faheem O, Thompson PD. The genetics of statin-induced myopathy. Atherosclerosis. 2010;210:337–43.

Giacomini KM, Huang SM, Tweedie DJ, et al. Membrane transporters in drug development. Nat Rev Drug Discov. 2010;9:215–36.

Giacomini KM, Balimane PV, Cho SK, et al. International Transporter Consortium commentary on clinically important transporter polymorphisms. Clin Pharmacol Ther. 2013;94:23–6.

Ingelman-Sundberg M, Sim SC. Pharmacogenetic biomarkers as tools for improved drug therapy; emphasis on the cytochrome P450 system. Biochem Biophys Res Commun. 2010;396:90–4.

Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacol Ther. 2007;116:496–526.

Johansson I, Ingelman-Sundberg M. CNVs of human genes and their implication in pharmacogenetics. Cytogenet Genome Res. 2008;123:195–204.

Jonas DE, McLeod HL. Genetic and clinical factors relating to warfarin dosing. Trends Pharmacol Sci. 2009;30:375–86.

Kazui M, Nishiya Y, Ishizuka T, et al. Identification of the human cytochrome P450 enzymes involved in the two oxidative steps in the bioactivation of clopidogrel to its pharmacologically active metabolite. Drug Metab Dispos. 2010;38:92–9.

Keskitalo JE, Zolk O, Fromm MF, Kurkinen KJ, Neuvonen PJ, Niemi M. ABCG2 polymorphism markedly affects the pharmacokinetics of atorvastatin and rosuvastatin. Clin Pharmacol Ther. 2009;86:197–203.

Kirchheiner J, Brockmoller J. Clinical consequences of cytochrome P450 2C9 polymorphisms. Clin Pharmacol Ther. 2005;77:1–16.

Kirchheiner J, Schmidt H, Tzvetkov M, et al. Pharmacokinetics of codeine and its metabolite morphine in ultra-rapid metabolizers due to CYP2D6 duplication. Pharmacogenomics J. 2007;7:257–65.

Kiyotani K, Mushiroda T, Sasa M, et al. Impact of CYP2D6*10 on recurrence-free survival in breast cancer patients receiving adjuvant tamoxifen therapy. Cancer Sci. 2008;99:995–9.

Kiyotani K, Mushiroda T, Imamura CK, et al. Significant effect of polymorphisms in CYP2D6 and ABCC2 on clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients. J Clin Oncol. 2010;28:1287–93.

Kiyotani K, Mushiroda T, Zembutsu H, Nakamura Y. Important and critical scientific aspects in pharmacogenomics analysis: lessons from controversial results of tamoxifen and CYP2D6 studies. J Hum Genet. 2013;58:327–33.

Link E, Parish S, Armitage J, et al. SLCO1B1 variants and statin-induced myopathy – a genomewide study. N Engl J Med. 2008;359:789–99.

Maliepaard M, Nofziger C, Papaluca M, et al. Pharmacogenetics in the evaluation of new drugs: a multiregional regulatory perspective. Nat Rev Drug Discov. 2013;12:103–15.

Mega JL, Hochholzer W, Frelinger III AL, et al. Dosing clopidogrel based on CYP2C19 genotype and the effect on platelet reactivity in patients with stable cardiovascular disease. JAMA. 2011;306:2221–8.

Niemi M, Schaeffeler E, Lang T, et al. High plasma pravastatin concentrations are associated with single nucleotide polymorphisms and haplotypes of organic anion transporting polypeptide-C (OATP-C, SLCO1B1). Pharmacogenetics. 2004;14:429–40.

Niemi M, Pasanen MK, Neuvonen PJ. Organic anion transporting polypeptide 1B1: a genetically polymorphic transporter of major importance for hepatic drug uptake. Pharmacol Rev. 2011;63:157–81.

Nies AT, Niemi M, Burk O, et al. Genetics is a major determinant of expression of the human hepatic uptake transporter OATP1B1, but not of OATP1B3 and OATP2B1. Genome Med. 2013;5:1.

Nishizato Y, Ieiri I, Suzuki H, et al. Polymorphisms of OATP-C (SLC21A6) and OAT3 (SLC22A8) genes: consequences for pravastatin pharmacokinetics. Clin Pharmacol Ther. 2003;73:554–65.

Paul E, End-Rodrigues T, Thylen P. Bergman U [Adverse drug reactions a common cause of hospitalization of the elderly. A clinical retrospective study]. Lakartidningen. 2008;105:2338–42.

Polgar O, Robey RW, Bates SE. ABCG2: structure, function and role in drug response. Expert Opin Drug Metab Toxicol. 2008;4:1–15.

Sanderson S, Emery J, Higgins J. CYP2C9 gene variants, drug dose, and bleeding risk in warfarin-treated patients: a HuGEnet systematic review and meta-analysis. Genet Med. 2005;7:97–104.

Schulz C, Boeck S, Heinemann V, Stemmler HJ. UGT1A1 genotyping: a predictor of irinotecan-associated side effects and drug efficacy? Anticancer Drugs. 2009;20:867–79.

Sharma P, Butters CJ, Smith V, Elsby R, Surry D. Prediction of the in vivo OATP1B1-mediated drug-drug interaction potential of an investigational drug against a range of statins. Eur J Pharm Sci. 2012;47:244–55.

Shitara Y, Horie T, Sugiyama Y. Transporters as a determinant of drug clearance and tissue distribution. Eur J Pharm Sci. 2006;27:425–46.

Sim SC, Ingelman-Sundberg M. Pharmacogenomic biomarkers: new tools in current and future drug therapy. Trends Pharmacol Sci. 2011;32:72–81.

Simon T, Bhatt DL, Bergougnan L, et al. Genetic polymorphisms and the impact of a higher clopidogrel dose regimen on active metabolite exposure and antiplatelet response in healthy subjects. Clin Pharmacol Ther. 2011;90:287–95.

Singh MS, Francis PA, Michael M. Tamoxifen, cytochrome P450 genes and breast cancer clinical outcomes. Breast. 2011;20:111–18.

Stearns V, Johnson MD, Rae JM, et al. Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. J Natl Cancer Inst. 2003;95:1758–64.

Stingl JC, Bartels H, Viviani R, Lehmann ML, Brockmoller J. Relevance of UDP-glucuronosyltransferase polymorphisms for drug dosing: a quantitative systematic review. Pharmacol Ther. 2014;141:92–116.

Tirona RG, Leake BF, Merino G, Kim RB. Polymorphisms in OATP-C: identification of multiple allelic variants associated with altered transport activity among European- and African-Americans. J Biol Chem. 2001;276:35669–75.

Voora D, Shah SH, Spasojevic I, et al. The SLCO1B1*5 genetic variant is associated with statin-induced side effects. J Am Coll Cardiol. 2009;54:1609–16.

Wienkers LC, Heath TG. Predicting in vivo drug interactions from in vitro drug discovery data. Nat Rev Drug Discov. 2005;4:825–33.

Wilke RA, Ramsey LB, Johnson SG, et al. The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy. Clin Pharmacol Ther. 2012;92:112–17.

Yang J, Chen Y, Li X, et al. Influence of CYP2C9 and VKORC1 genotypes on the risk of hemorrhagic complications in warfarin-treated patients: a systematic review and meta-analysis. Int J Cardiol. 2013;168:4234–43.

Yin T, Miyata T. Warfarin dose and the pharmacogenomics of CYP2C9 and VKO. Thromb Res. 2007;120:1–10.

Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013;138:103–41.

Chapter 12
The Role of Proteomics in the Development of Personalized Medicine, Diagnostic Methods and Large Scale Biobanking

Johan Malm and György Marko-Varga

Abstract The current field of biobanking is emerging and expanding very fast with both private and governmental support. We present here based on our own developments recommendations and outlines for standardization in biobanking processes in clinical studies.

The biobank processes are important for the use of samples in developing assays. These measurements are very important in order to document states of health and disease. The data generation and diagnostic indications are beneficial for academic research, commercial healthcare, drug development industry and government regulating agencies.

There is a need for an improved awareness within proteomic and genomic communities regarding the basic concepts of collecting, storing and utilizing clinical samples. In this respect, the aspects of quality control and sample suitability for analysis need to be documented and validated to ensure data integrity and establish contexts for interpretation of results. The current presentation and outline is of major relevance to the proteomic and genomic fields. The standardization aspects of biobanking and the requirements that are needed to run future clinical studies that will benefit the patients where OMICS science will play a major role. A global view of the field is given where best practice and conventional acceptances are presented along with ongoing large-scale biobanking projects.

Keywords Biobank • Regulatory • Ethics • LIMS • Proteomics

J. Malm (✉)
Section for Clinical Chemistry, Department of Laboratory Medicine, Malmö,
Lund University, 205 02 Malmö, Sweden
e-mail: johan.malm@med.lu.se

G. Marko-Varga
Clinical Protein Science and Imaging, Biomedical Center, Biomedical Engineering,
Lund University, BMC D13, 221 84 Lund, Sweden

Center of Excellence in Biological and Medical Mass Spectrometry,
Biomedical Center D13, Lund University, 221 84 Lund, Sweden

First Department of Surgery, Tokyo Medical University,
6-7-1 Nishishinjiku Shinjuku-ku, Tokyo 160-0023, Japan
e-mail: gyorgy.marko-varga@bme.lth.se

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_12

Abbreviations

SNPs	Single Nucleotide polymorphisms
IRB	Institutional Review Board
LIMS	Laboratory Intelligence Management Systems
BBMRI	The Biobanking and Biomolecular Resources Research Infrastructure
EC	European Commission
EU	European Union

12.1 Introduction

Modern healthcare is looking for ways to treat patients that are more cost effective without losing out on the care that patients demand today. Large scale biobank repositories constitute a key component in improving diagnostic methods (Riegman et al. 2008; Khleif et al. 2010; http://www.informatics-review.com/wiki/index.php/Biobanking_Definition). Today, many millions of clinical samples are collected every day for use in diagnostic tests that support clinical decision making. Worldwide, it is estimated that over one billion clinical samples are assembled into so called biobanks, also known as biospecimen resources, and stored (Lasso 2010).

A biobank can be defined as a storage facility where long-term storage of human samples that are traceable to a specific person and linked to personal data. As a new trend in healthcare activities that relate to epigenetics and epidemiology, population-based research biobanks will also collect environmental and lifestyle information and generate data that enables large scale meta data analyses. Not only the health care system but also drug discovery and drug development within the pharmaceutical industry is heavily dependent on biobank resources for future work. In this respect the new generation of Personalized Medicine (PM) is an industrial objective with top priority (Hewitt 2011, Hamburg and Collins 2010, Marko-Varga et al. 2007). Time magazine recently named Biobanking one of the “10 Ideas Changing the World Right Now” (Park 2009).

Healthcare organizations such as hospitals and clinical institutions, both private and public, are responsible for most biobank samples and use them in various screening programs, in diagnostics and in quality improvement processes. In recent years, the ethical debate and best practice within the biobank field has focused on the patients’ rights (Lasso 2010; Simeon-Dubach and Perren 2011). An area that has been given special attention was the FDA Critical Path Initiative. This was directed towards the development of better evaluation tools like e.g. biomarkers and in subsequent recommendations from the AACR–FDA–NCI (AACR, American Association for Cancer Research; NCI, US National Cancer Institute) Cancer Biomarkers Collaborative group, emphasis was given to the need for improved biobanking services and biospecimen quality control (Vaught 2006).

The Nordic countries have a long tradition in biobanking, such as the Twin study with 80,000 twin pairs (http://ki.se/ki/jsp/polopoly.jsp?d=12484&a=26264&l=sv), the LifeGene study (www.lifegene.se), and National studies in Finland; http://www.p3gconsortium.org/about.cfm and http://www.decode.com/. In the Norwegian "Biohealth" program the Institute for Public Health (NIPH) coordinates large population cohorts including more than 400,000 subjects. The Norwegian Mother and Child cohort study includes 210,000 subjects, and the ultimate goal is to include 500,000 individuals. Denmark also has several large research cohorts, e.g. the National Birth Cohort "Better health for mother and child" with blood from 100,000 pregnant women, and the Nutrition, Cancer and Health biobank with around 60,000 subjects. In Greenland, the national Biobank has blood samples from close to 20 % of the population. The European Union initiative "BBMRI", The Biobanking and Biomolecular Resources Research Infrastructure, (http://bbmri.eu/sv) has now grown into a 54-member consortium with more than 225 associated organisations (largely biobanks) from over 30 countries, making it one of the largest research infrastructure projects in Europe. BBMRI was one of the first projects entering the European Research Infrastructure preparatory phase of the ESFRI (European Strategy Forum on Research Infrastructures) roadmap funded by the European Commission (EC).

Clinical samples in biobanks have become an important asset and are now used in health care. Most of these samples are categorized as containing: tissues, cells/cell lines, genomic material (DNA), blood or blood-plasma or urine. Fractionated blood samples and paraffin blocks of tissues constitute the majority of samples. In Sweden, the vast majority of these samples can be identified within the biobank registry at the national board of health and welfare http://www.biobanks.se/.

Rapid progress in hospitals utilizing genomics and proteomics research fields is expected (Malm et al. 2013; Marko-Varga 2013; Welinder et al. 2013; Marko-Varga et al. 2012). A number of patients will have conditions that do not arise from a specific and single pathology, but rather is the result of a combination of factors, multifactorial diseases. High-throughput technologies that generate global expressions and analyses of biological systems are expected to allow better molecular understanding of heterogeneous and complex diseases. The outcome of breakthroughs in future research by unraveling the pathophysiological mechanisms of diseases will depend on the study of large sets of well-documented, epidemiologically supported, and clinically verified biological and molecular information and bio-samples from large cohorts of patients and healthy persons, that are made available through biobanks (Baker 2012). Clinical samples processed and archived in biobanks is a key resource for developing a disease understanding.

Currently the areas of strategic importance for current and future healthcare developments include: diagnostics developments, patient stratification, quality assurance, education, research drug discovery/development and clinical trials.

Diagnostic developments based on rapid technological progress within the field of mass spectrometry have made expectations high that some clinical tests that traditionally have been performed using immunoassays will be replaced by tests based on mass spectrometry. Expectations have also been high that new biomarkers, discovered using proteomics, will be introduced in clinical routine diagnostics and used for stratification of patients for specific treatment and for patient monitoring. The basis for this optimism is the lack of interference associated with mass spectrometry (compared to immunoassays), less volume used, the possibility to design multiplex assays, low reagent costs, and the fact that the 'exact' analyte is much better defined.

In spite of this optimism the introduction of mass spectrometry based assays in clinical medicine has been a slower process than expected by many experts, especially experts from the outside hospitals and hospital laboratories. Two major reasons can easily be identified – the complexity of the assays and the lack of studies on the clinical utility of MS-based assays, e.g. biomarkers.

Most assays run in clinical laboratories are commercial assays, analytically well validated, easy to perform with instruments that are more or less automated thus minimizing manual steps. The ease of use is important since skilled and experienced laboratory staff is hard to find in many countries. The quality of respective assay is usually monitored by e.g. using control samples and by participating in external quality assurance programs. The analytical requirements on a routine clinical assay are in general very high.

The economic constraints on healthcare have become harder as a consequence of the financial problems in many countries in recent years, governmental spending on healthcare is often being reduced. Although laboratory medicine constitutes only a small portion of health care costs laboratory leaders and hospital administrators are more reluctant to introduce new tests, at least outside university hospitals, unless the demand for the new test can be expected to be reasonably high.

In the past, economical considerations associated with the introduction of a new laboratory test have usually been made often only to a minor extent, and often focusing on how much the vendor charges the laboratory. In the future, it can be expected that more focus will be on the economic benefit in addition to clinical utility (Scott 2010).

In order for mass spectrometry based analyses to be accepted by the healthcare community there are a number of aspects that must be taken into consideration for the assay to be accepted, and used for diagnostic and therapy monitoring purposes. The requirements are basically the same as for today's routine analyses.

12.2 Biobanking – Regulatory Aspects

The ethical and legal regulations governing the use of biobanked samples are determined by public law in place at the location of sampling and at the site of analysis. These governing rules pertain to both academic and commercial use of the samples. Paramount to this point is the voluntary subject informed consent giving permission for the specific use of these samples and the approval of an institutional review board guaranteeing the safety of the subject in obtaining the sample and in

Table 12.1 Biobank repositories of clinical samples

Location	Management		Governance
Local	Single investigator		Institutional review board
National	Institutional		Informed consent
Multi-national	Multi-center		National law
Global	Commercial		International law
Focus	Stored biospecimen		Storage conditions
Single disease	RNA	Fluids	Short term −20 °C
Complex disease	DNA	Cells	Long term −80 °C
Inherited disease	Proteins	Tissue	Long term −260 °C
Environmental	Peptides	Organs	Room temperature
Rare diseases	Lipids	Body	
Population based studies	Metabolites		
Drug/clinical trials			

the use of these samples. Globally the ethical use of clinical samples is covered by the United Nations Universal Declaration of Human Rights and the Declaration of Helsinki by the World Medical Association. Throughout the European Community, both The Council of Europe and individual countries (Denmark, Estonia, Finland, France, Germany, Iceland, Norway, Sweden, and the United Kingdom) have enacted regulations governing biobanks. State enacted public law is also in place in Australia, Canada, and Latvia with legislation pending in many more (Table 12.1).

A major responsibility of the biobank establishments and organizations worldwide is to protect the donor against research risks. Documentation that protects donors is of greatest importance. Donors must be confident that the aims, objectives and delivered values of a donation are respected. Otherwise, in worst cases, clinical study initiatives with biobank establishments might be seriously compromised.

Personal patient data must have secure safety systems to protect against risk of being inappropriately utilized by third parties, such as insurance companies, employers, and others. This is needed, in order to guarantee, and provide patient confidentiality and data protection. Safe net mechanisms are documented and used in every day medical practice and also include ethics committee reviews of research projects requiring bio-specimens. This also includes the informed consent, that is a documented guarantee that will protect the interest of the patient with respect for autonomy (The Biobank investigation).

Overall, there are a number of laws on a national level within countries that regulate the use of biobanks. In Sweden for instance, these regulatory directives relate to the ethical considerations that need to be taken into consideration, establishing and using patient materials, and patient data (Ethic Review Act (SFS 2003:460)). In addition, there are several laws that regulate the build, and the usage, such as the Secrecy Act (SFS 1980:100), as well as the Biobank Act (SFS 2002:297), that directly relate to bio-repositories and biobanking. Other considerations that are of utmost importance, and regulated by laws, which are related to secrecy, and personal patient integrity protection, are the; the secrecy act (SFS 1980:100) as well as the personal data act (SFS 1998:204).

In this respect, The Biobank Act, has a dual role on a national level in Sweden. The aim is to protect donor integrity, while also promoting research on biobank samples. The National Board of Health and Welfare is the central government authority commissioned to implement the biobank Act. The code of conduct in this respect is that the board develops regulations and practical rules, such as the; SOSFS 2002:11, SOSFS 2004:2 and SOSFS 2006:19, that are enforced and will administer how to apply respective law.

The structure whereby Swedish biobanks are established and used is based on the biobank resources formed in Sweden by public or private health services, denamed: “primary biobanks”, or to biobanks formed by using samples from a primary biobank, denamed: “secondary biobanks”. It is important to recognize that this Act will not apply to biobanks that have been formed and assembled by any organization other than a health care provider, such as, healthcare consortia, that will enroll participants in for instance population based studies where the sampling is made in public places or in dedicated facilities, other than hospitals and healthcare institutions. Other examples are study initiatives that are conducted by pharmaceutical companies. However, in all cases, these study initiatives and biobank establishments must all be registered at the National Board of Health and Welfare. Further, biobank samples that are confined within a secondary biobank are not allowed to be distributed to a third party. The biobank Act also specifically highlights that the informed consent from the donor will be required in order to be able to store, and make use of the human samples.

Clinical information and test results, such as Genomics and Proteomics sequences, stored in the hospital databases, have to be fully integrated behind a safety fire wall, that provides fully identified patient data, including personal, clinical and laboratory information (Vaught 2006; Malm et al. 2013; Marko-Varga 2013). Biobanks are generally under the authority of an institutional review board.

Currently, within the European Union countries, which have approved, and are members of the BBMRI, as well as at an international level, more work is needed in harmonizing the best practice and operative structures of biobanks. To meet these goals, there are a number of collaborative efforts that drive these developments, such as a large number of BBMRI initiatives. Synchronize and coordinating international collaboration is the way forward to reach a unity for global use of patient material, improving our healthcare service system.

12.3 Patient Donors and Sample Integrity

Scandinavian countries, including Sweden, has a historical record and an advantage, given its many population based registries and databases. As an example of the structure and organization in place in the healthcare sector, at a national level, hospital registries in Sweden maintain detailed registers for epidemiological analyses, providing a treasured resource for biobank research. In addition, national hospital systems maintain valuable quality registries in order to follow up on patient treatments that allows to treat with the disease history at hand, including demographic databases. Additional registries, that bring high value and facilitate optimal treatment, as well as high quality biobank research are:

- Prescribed Drug Registry, all prescriptions since 2005
- Cancer Registry, all cancer cases since 1958
- Cause of Death Registry, all underlying causes since 1952
- Medical Birth Registry, all births since 1973
- Hospital Discharge Registry, all diagnoses and medical treatments since 1961

One challenge right now is to establish an efficient linkage in-between these databases and biobank databases. Laboratory Intelligence Management Systems (LIMS) are necessary when introducing automated processing with large study sample numbers. The LIMS system will keep track of each individual sample, as well as track the time and modular handling that has taken place with respective biobank sample. This is important, as the unidentified samples in the biobank freezers can be identified by the barcode of the sample tube. These barcode identifiers are electronically registered and followed over time, during the entire lifetime of the sample. In this way, barcoded samples can be scanned and tracked from collection to analysis.

As we are utilizing identifiers of each sample, its origin will correspond to a given study that is related to a given patient and sample type. The entire work flow and process handling is run under electronic surveillance, that ensures the sample integrity of each biobank unit. The LIMS systems are also a vital tool in the expansion of international large scale studies, as it allows a build of multi-study experimental screenings. This integration that is built with LIMS capabilities, also favours the willingness of patients to donate bio-fluids, tissues and share disease experience by filling out questionnaires. There is an additional gain by utilising fully automated biobank systems with LIMS and electronic surveillance, which relate to an ethical strengthening and motivation. The new generation of biobanking with these efficient processing technologies also provides an element of confidential openness, where the overview of patient safety and sample usage can be documented.

12.4 Proteomics in Clinical Diagnostics – Analytical Aspects

Proteomics has revolutionized protein identification and allowed for resolution of complex proteomes in a couple of days and the design of multiplex quantitative assays and it is believed that many immunoassays will be replaced by more specific MS based assays within the next decade or so. However, mass spectrometry, like any other analytical methodology, has its limitations. For using mass spectrometry in a clinical setting the most important limitation is probably sensitivity. Sample preparation for quantitative mass spectrometry often includes enzymatic digestion, e.g. with trypsin, and as a result the initially complex mixture, e.g. serum, becomes even more complex, composed of millions of peptides and thus competition for ionization. Even after an initial chromatographic separation step this results in suppression of full ionization and substantial loss of sensitivity since the peptide of interest is ionized to a much lower extent than had it been present in pure form. In general, immunoassays are often one to two magnitudes more sensitive than the corresponding mass spectrometry based method. Enrichment steps can improve the sensitivity but these techniques often compromise throughput. It can thus be expected that more time is needed for improvements in both sample pretreatment and instrumentation before mass spectrometry based methods will replace immunoassays to any larger extent.

For an assay based on proteomics to be introduced in a clinical setting the molecule(s) of interest must first be selected, validated and standard operating procedures (SOPs) established. The human proteome has a large dynamic range of protein concentrations and the proteome is extremely complex. Depending on their origin proteins may exist in multiple forms due to post-translational modifications. Degradation processes, in vivo as well as in vitro, may result in yet additional molecular forms with varying biochemical characteristics. Some of these peptides may be formed in vitro and be disease-specific whereas others are the result of in vitro degradation that may depend on preanalytical variables like processing temperature, type of sample container (glass, plastic, type of anticoagulant, gel-based separator), clotting time, mode of specimen collection (needle bore size, patient posture), time between venipuncture and separation of serum/plasma and time of storage and freeze-thaw cycles. Also factors concerning patient preparation, e.g. fasting and time of sample collection, may affect the result of subsequent analyses.

In general, more attention should be paid to the preanalytical steps – sample collection, transportation and preanalytical processing. These measurements are very important also in the discovery and validation phases when experience of the analyte(s) is limited. The importance of the preanalytical phase for many clinical chemistry and hematological parameters is well known and it can be expected that proteomics based analytes, occurring at low concentrations in complex mixtures and measured with complex and sensitive methodology are likely to be influenced by preanalytical handling. Since it is, at least in most cases not possible to foresee the result of a certain preanalytical procedure standardization is of greatest importance to obtain comparable and reproducible results both within the laboratory and between different laboratories (Apweiler 2009).

The biomarker(s) to be validated must be clearly defined, not only with regard to amino acid sequence but also any post-translational modifications (Bozovic and Kulasingam 2013). When determining which parameters to evaluate it is important to consider both the intended use of the assay and the requirements from regulatory bodies. In most cases, the same parameters as for routine clinical chemistry assays should be characterized and documented. It is also important to keep in mind that transfer of a mass spectrometry based analysis to a clinical routine laboratory, e.g. in a hospital, may not always be the best option. The assay may be too complex, the throughput too low and the assay not robust enough. An alternate method should always be considered, in some cases an immunoassay is a better option.

A prerequisite for a clinically useful quantitative assay is a pure reference standard and an internal standard for each analyte to be measured. A pure standard allows for analysis of absolute concentration and selection of ions and thus make certain that different assay parameters, e.g. retention time, are representative of the analyte. In cases when reference standards are not commercially available and custom made compounds is the only alternative it is important to use material of high purity and verify it against proficiency testing samples or other assay techniques. As for other clinical assay methods it is an advantage to have reference standards at concentrations close to the cut off point, the upper and lower reference limits or the decision points. In order to avoid disturbing matrix effects the standard material should be used in a matrix similar to that of the samples to be analyzed.

Clinical proteomics is often based on analysis of serum, a very complex mixture of small and large molecules of various nature present at concentrations differing in several orders of magnitude. Prior to the mass spectrometric analysis a preanalytical sample preparation is often necessary, a procedure that may impact the amount and nature of the analyte. In order to ensure that the signal measured by the instrument represents the analyte and that the impact of the preanalytical procedure is compensated for an internal standard is often used, particularly for quantitative mass spectrometry.

Internal standards can be composed of structural analogues of the analyte but in most cases the internal standard is an analogue of the analyte. Commercially available standards often contain isotopes of hydrogen (deuterium) or carbon. These compounds usually have biophysical properties more or less identical to those of the analyte. The internal standard must not be present in the sample at amounts higher than a few percent of the analyte, otherwise the quantification can be impacted.

Once the optimization of liquid chromatography and mass spectrometry parameters has been performed (not discussed here) the assay intended for clinical used has to be validated. This should be done prior to implementation of the method in clinical routine. Many clinical laboratories are accredited by national regulatory agencies that require well documented assay characteristics. These agencies usually perform annual formal evaluations of clinical laboratories and in order to satisfy healthcare professionals and patients and get renewed accreditation clinical proteomic based assays must be as robust and reliable as any other assay. Clinicians often do not understand the difference between a ‘simple’ test and a very complex test – they expect all the test results from e.g. the hospital laboratory to be of the same, high, quality.

When the method validation is complete the assay characteristics should be summarized in a document that the laboratory can present to customers and regulatory agencies. In most cases the validation includes evaluation of precision, sensitivity, specificity, linearity, carryover and the impact of different matrices and interfering substances. The acceptance criteria should primarily be based on medical requirements, including biological variation, but also assay characteristics from alternative methods, requirements from regulatory agencies and economical aspects need to be considered.

The most basic evaluation of assay sensitivity is to analyze blank samples, e.g. urine or serum. A low background indicates that a high signal to noise ratio may be expected. Several samples, from both healthy individuals and patients, should be studied. Both patients on medications that can be expected to interfere (i.e. patients using structurally similar compounds) and patient samples that are known to be problematic, e.g. icteric, lipemic and hemolytic samples should be studied. If serum is the preferred sample type it is often wise to study also plasma since this will be sent to the laboratory, by mistake, in some cases.

Since interfering substances may co-elute in the chromatography the specificity can also be evaluated by studying several transitions for respective analyte and calculating their relative ratios. For reasons mentioned above several samples should be studied.

A complement to using samples from patients on different medications is to spike blank samples and samples containing the analyte, with compounds related to the analyte, compounds that could potentially interfere in the assay.

The qualitative impact of different matrices is often studied by the post-column infusion spike technique whereas quantitative matrix effects are studied by the post-extraction spike technique (Taylor 2005; Chambers et al. 2007). Spiking experiments are also performed for recovery studies.

Once the proteomics based assay is introduced by the clinical laboratory most clinicians will not question the specificity and selectivity of the assay since most other routine assays are not associated with specificity or selectivity problems. In contrast, the accuracy and the precision of a routine assay are more often discussed since test results for an individual patient may come from different laboratories using different instruments which sometimes give rise to discrepant results. Accuracy and precision are also often discussed when test results are used in clinical studies.

Accuracy refers to the bias of a method, the difference between the mean value of the measured analyte and the true value – often a value from a reference laboratory or analysis of a sample with an assigned value. The accuracy should be measured at concentrations close to reference limits or decision points – three different levels are often used. Also the precision should be measured at reference limits or at decision points and include both within-run and between-day (or between-run) precision. In order to include any matrix effect the precision should be determined using a sample with the same matrix as patient samples. The precision is often overestimated when it is calculated based on assay results of the analyte spiked into e.g. a buffer.

Two more clinically important variables that should be determined are the limit of detection (LOD), i.e. the lowest concentration that can be distinguished from a sample that does not contain the analyte, and the limit of quantification (LOQ), i.e. the lowest concentration that can be measured with acceptable precision and accuracy. Depending on the clinical assay requirements the LOQ is equal or higher than the LOD.

12.5 Concluding Remarks

The key area for progress using archived samples is within the field of new biomarkers and designing new diagnostic tools based on new biomarkers. Both industry and academia are investing and searching for approaches to improve on the discovery successes where new technology plays a central role. Here, the ENCODE initiative has over the years produced an extensive DNA-sequence resource that is being utilized by the Proteomics Community.

The human genome sequencing platforms including the latest generation of deep-sequencing platforms, allows us to integrate new data with genetic risk factors. These risk factors can be correlated with demographic and lifestyle data collected via modern communication technologies. The technical prerequisites now exist to merge large volumes of molecular genetic data obtained by using new high-throughput DNA analysis platforms with clinical, epidemiological and national health registry data. Together with other global datasets from transcriptomics and proteomics analyses of biobank samples, these provide completely new opportunities to develop new cures and diagnostics that address common multifactor diseases of different backgrounds (The HapMap project). Furthermore, scientists have found that people vary not only by single nucleotide polymorphisms (SNPs), but that some people differ in large blocks of DNA, which are deleted or inserted. Until recently, the major focus was to determine how genetic polymorphisms influence protein structure and function (coding SNPs).

However, approaches with global analysis utilising expression microarrays have demonstrated that small differences in an individual's DNA may affect disease risk by altering the regulation of gene expression, thus modifying the amount of protein produced in respective cell (regulatory SNPs). These disease-associated polymorphisms provide a guide to potential molecular alteration. The consequence in some cases, but not all, is a shift in protein sequences that is related to the disease mechanisms and progression. As we learn more about how these polymorphisms change the function of genes, proteins, cells and organs, there is an opportunity to link these to make predictions in DNA sequence alteration in patient cohorts.

The process of introducing proteomics into clinical routine diagnostics has been slower than expected. From discussions with clinicians and laboratorians two important reasons can be identified, lack of clinically well validated biomarkers for risk stratification, early detection, prediction, and disease prognosis and lack of commercially available, robust assays that can be run also in non-highly-specialized hospital labs.

To improve the situation a collaboration between research labs, clinical labs, industry and clinicians is necessary. The economic pressure on hospitals has increased substantially in recent years and more focus than ever is on the economic value of a new test. Can it reduce the hospital time for the patient? Is it substantially better than the presently available tests? Does it answer any questions that cannot be answered by other diagnostic tools? The two last questions are critical – a new test without a clearly added value is very hard to introduce in today's routine patient care. Probably much harder than just a decade ago.

The introduction of proteomics into clinical medicine requires robust and clinically validated methods for application on large sample sets. A general acceptance of these methods by clinicians depends on studies focused on quantitating proteins in specific populations of patients showing that proteomics based assays are clinically and economically justified.

Acknowledgement and Funding This work was supported by grants from Mrs. Berta Kamprad Foundation, Swedish Academy of Pharmaceutical Sciences, the Swedish Foundation for Strategic Research, Vinnova, Fundacion Federico SA, Ingabritt & Arne Lundbergs forskningsstiftelse and the Crafoord Foundation.

References

Apweiler R. Approaching clinical proteomics: current state and future fields of application in fluid proteomics. Clin Chem Lab Med. 2009;47(6):724–44.

Baker M. Biorepositories: building better biobanks. Nature. 2012;486:141–6.

Bozovic A, Kulasingam V. Quantitative mass spectrometry-based assay development and validation: from small molecules to proteins. Clin Biochem. 2013;46:444–55.

Chambers E, Wagrowski-Diehl DM, Lu Z, Mazzeo JR. Systematic and comprehensive strategy for reducing matrix effects in LC/MS/MS analyses. J Chromatogr. 2007;852(1–2):22–34.

Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. 2010;363:301–4.

Hewitt RE. Biobanking: the foundation of personalized medicine. Curr Opin Oncol. 2011;23(1):112–19.

Khleif SN, Doroshow JH, Hait WN. AACR–FDA–NCI cancer biomarkers collaborative consensus report: advancing the use of biomarkers in cancer drug development. Clin Cancer Res. 2010;16(13):3299–318.

Lasso RO. The ethics of research biobanking. JAMA. 2010;304(8):908–10.

Malm J, Fehniger TE, Danmyr P, Vegvari A, Welinder C, Lindberg H, Appelqvist R, Sjödin K, Wieslander E, Laurell T, Hober S, Berven FS, Fenyö D, Wang X, Andrén PE, Edula G, Carlsohn E, Fuentes M, Nilsson CL, Dahlbäck M, Rezeli M, Erlinge D, Marko-Varga G. Developments in biobanking workflow standardization providing sample integrity and stability. J Proteomics. 2013;95:38–45.

Marko-Varga G. BioBanking as the central tool for translational medicine CTM issue 2013. Clin Trans Med. 2013;2(4):1–4.

Marko-Varga G, Ogiwara A, Nishimura T, Kawamura T, Fujii K, Kawakami T, Kyono Y, Tu HK, Anyoji H, Kanazawa M, Akimoto S, Hirano T, Tsuboi M, Nishio K, Hada S, Jiang H, Fukuoka M, Nakata K, Nishiwaki Y, Kunito H, Peers IS, Harbron CG, South MC, Higenbottam T, Nyberg F, Kudoh S, Kato H. Personalized medicine and proteomics: lessons from non-small cell lung cancer. J Proteome Res. 2007;6(8):2925–35.

Marko-Varga G, Vegvari A, Welinder C, Lindberg H, Rezeli M, Edula G, Svensson KJ, Belting M, Laurell T, Fehniger TE. Standardization and utilization of biobank resources in clinical protein science with examples of emerging applications. J Proteome Res. 2012;11:5124–34.

Park A. Biobanks. "Ten ideas changing the world right now." TIME Magazine. 2009, March 9 2009.

Riegman PHJ, Morente MM, Betsou F, De Blasio P, Geary P, Marble Arch Int Working G. Biobanking for better healthcare. Mol Oncol. 2008;2(3):213–22.

Scott MG. When do new biomarkers make sense? Scand J Clin Lab Invest. 2010;70 suppl 242:90–5.

Simeon-Dubach D, Perren A. Better provenance for biobank samples. Nature. 2011;475(7357):454–5.

Taylor PJ. Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography-electrospray-tandem mass spectrometry. Clin Biochem. 2005;38(4):90–5.

Vaught JB. Biorepository and biospecimen science: a new focus for CEBP. Cancer Epidemiol Biomarkers Prev. 2006;15:1572–3.

Welinder C, Jonsson G, Ingvar C, Lundgren L, Olsson H, Breslin T, Végvári A, Laurell T, Rezeli M, Jansson B, Baldetorp B, Marko-Varga G. Establishing a Southern Swedish Malignant Melanoma OMICS and biobank clinical capability. Clin Transplant Med. 2013;2(1):1326–7.

Chapter 13
Imaging Techniques in Proteomics Research

Devipriya Subramaniyam and Goutham Edula

Abstract Imaging has been used for several decades for the visual representation of cellular and molecular processes of living organisms in a two or three dimensional fashion. Several imaging techniques like X-rays, immunohistochemistry, nuclear magnetic resonance, cryo-electron microscopes, positron emission tomography, green fluorescent protein labelling etc., have been developed and used to understand the structure and function of biological compounds (Saito et al. 2012). These techniques have aided in study of the structure and function of several proteins thereby facilitating the understanding of various human diseases. As advancement to the available imaging methods, a new technique called the Mass Spectrometry Imaging (MSI) or the Imaging Mass Spectrometry (IMS) has gained momentum in the recent past and has enabled the analysis of spatial distribution of biomolecules like peptides, metabolites, proteins etc., as well as pharmaceutical compounds based on their molecular masses. In this chapter we focus on the various strategies used in MSI experiments, the types of compounds analysed and the methodology and analyses used by MSI.

Keywords Imaging • MALDI • MSI • Pharmaceuticals • Lipids • Neurotransmitters • Peptides • Proteins

13.1 Introduction

Imaging has been used for several decades for the visual representation of cellular and molecular processes of living organisms in a two or three dimensional fashion. Several imaging techniques like X-rays, immunohistochemistry, nuclear magnetic resonance, cryo-electron microscopes, positron emission tomography, green fluorescent protein labelling etc., have been developed and used to understand the structure and function of biological compounds (Saito et al. 2012). These techniques

D. Subramaniyam • G. Edula (✉)
Imaging Sciences Clinnovo Research Labs, Plot No: 4, Survey No: 11/2, KhanaMet,
Hitech City, Hyderabad, India
e-mail: goutham.edula@clinnovo.com

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*,
Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8_13

have aided in study of the structure and function of several proteins thereby facilitating the understanding of various human diseases. As advancement to the available imaging methods, a new technique called the Mass Spectrometry Imaging (MSI) or the Imaging Mass Spectrometry (IMS) has gained momentum in the recent past and has enabled the analysis of spatial distribution of biomolecules like peptides, metabolites, proteins etc., as well as pharmaceutical compounds based on their molecular masses. In this chapter we focus on the various strategies used in MSI experiments, the types of compounds analysed and the methodology and analyses used by MSI.

13.2 Mass Spectrometry Imaging (MSI)

Generally, proteomic studies require homogenization of a sample as well as long, often tedious and time-consuming extraction, purification, and separation steps before molecular characterization is performed, mostly without preserving anatomical information to correlate location in the tissue with expression profiles of particular peptides and proteins (Aebersold and Goodlett 2001). MSI overcomes these hurdles as it combines the advantages of Mass Spectrometry (MS) and microscopic imaging in a single experiment to assay molecular profiles directly from frozen or preserved tissue slices, reducing the number of preparative steps, while preserving topographical information about molecular distributions and localization (Chaurand et al. 1999; Fournier et al. 2003). Also, MSI is essentially a label free technique, thus making it possible to detect and characterize various known and unknown analytes without having to develop special labels. MSI uses the principles of Mass Spectrometry where the compound to be analysed is first ionized and the mass-to-charge (m/z) ratio of the resulting ions is determined, giving an indication of each analyte's atomic composition. Hundreds of molecules can be detected and identified simultaneously at μm or even sub-μm spatial resolutions in complex biological samples (Hillenkamp et al. 1975). Due to its high sensitivity, high multiplexing capabilities and mass accuracy, today MSI is preferred over other methods for the detection, identification and structural characterisation of proteins and peptides (Aebersold and Goodlett 2001). MSI have also been used to characterize tissues, including human gliomas and lung cancers, as well as tumor response to specific therapeutics, suggesting the use of proteomic information in assessing disease progression as well as predicting patient response to specific treatments (Chaurand et al. 2005). Thus, MSI has emerged as the most powerful tool in Proteomics research for mapping of potentially all proteins and peptides on tissue samples and for the identification and determination of tissue specific disease markers.

Over the last decade the MSI technology has seen tremendous developments allowing the analysis of a wide variety of compounds including inorganic elementals, metabolites, lipids, peptides, proteins and xenobiotics with spatial resolutions from micrometer to nanometer scales. The kind of information retrieved from an MSI experiment can vary depending on various factors such as, ionization procedures, sample preparation methods, spatial resolution and the speed of the technique.

Several approaches have been developed for performing MSI till date and they are classified according to the method of sample ionisation as follows:

- Matrix-Assisted Laser Desorption Ionization (MALDI)
- Secondary ion mass spectrometry (SIMS)
- Desorption Electro Spray Ionization (DESI)
- Rapid Evaporative Ionization Mass Spectrometry (REIMS)

13.3 Matrix-Assisted Laser Desorption Ionization (MALDI)-MSI

MALDI MSI provides a means for identifying the spatial distribution of different kinds of molecules on tissue sections including peptides and metabolites as well as to measure their abundance. MALDI MSI technique for the imaging of biological tissue samples was first described by (Caprioli et al. 1997). Since then, the method has undergone several modifications to improve the sensitivity, image resolution, sample preparation, date acquisition and analysis speed etc., and is constantly under development in several laboratories seeking to design new instruments and to improve the various steps involved in the technique as well as identifying new applications, such as studies of tissue localization of drugs, biomarker discovery, or understanding molecular mechanisms.

In a typical 2D MALDI MSI measurement, microtome tissue sections from fresh organ or biopsy are transferred and fixed to a target plate. Sections are then covered with a specific matrix compatible with the specific molecular species targeted for detection (typically proteins/peptides or metabolites) by microspraying, or microspotting, or by sublimation. For peptides/proteins, very intense signals are obtained with a-cyano-4-hydroxycinnamic acid (a-CHCA) as a matrix. Once the sections are covered with a-CHCA (or another matrix), they are dried in a vacuum desiccator and introduced into the vacuum inlet of a mass spectrometer. During this process, matrix solvents extract analytes from the tissue, quench endogenous proteolytic enzyme activity and eventually evaporate, leaving the analyte-doped matrix crystals. Then, a laser beam is rastered across the entire tissue surface over predefined 2D grids. A mass spectrum is acquired from ions desorbed and ionized from each irradiation surface spot and is recorded corresponding to each grid coordinate. From the intensity of a designated m/z ion detected in each spectrum, a 2D ion density map can be constructed, showing its relative abundance in specific regions (Chaurand et al. 2004) (Fig. 13.1). A deeper, volumetric view of analytes is subsequently achieved by serial cutting. A stack of 2D images from serially cut tissue sections with appropriate spacing is obtained and stitched together with 3D processing and visualization software to construct a 3D volumetric model.

Proteome with mid to low molecular weights can be studied effectively by direct analysis of tissue that serves both biological and clinical interest with the help of MALDI – MSI. This technique can analyse intact tissue without homogenization and separation steps and retain the spatial distribution of molecules within the

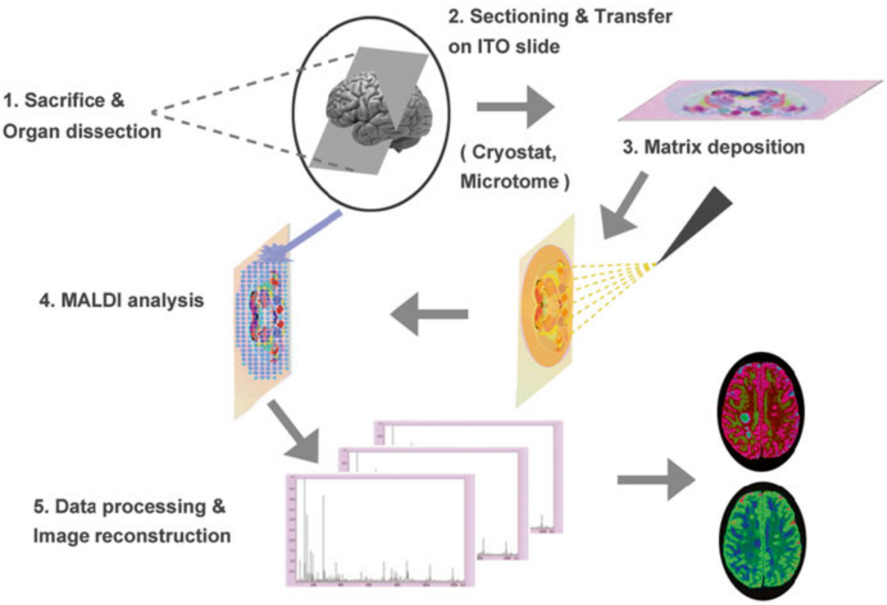


Fig. 13.1 Schematic representation of the experimental workflow of MALDI MSI

tissue. Pixel size in this technique is typically on the order of 1 to 10 μm or so, thus it can achieve subcellular resolution (Seeley and Caprioli 2008). The primary use of this technique is profiling and acquiring images of peptides and proteins from tissues and cell samples and also in the analysis of smaller molecules like drugs and its metabolites. The use of imaging using MALDI for smaller molecules can be challenging mainly because the low mass range is disrupted by the clusters formed of matrix ions and mixed analyte-matrix. This is a limitation as it affects the detection of analyte ions of less than 750 Da (Dalton) (Caprioli et al. 1997; Fournier et al. 2003; Seeley et al. 2008). Also, the matrix, which serves to absorb the laser energy and transfer it to the sample itself, can be difficult to apply and produces an abundance of small molecular weight ions, which can obscure the metabolite region of the resulting spectra. Another major limitation of MALDI MSI is the requirement for sample preparation and the need to analyze specimens under vacuum, thus limiting the possibility to study live biological samples (Chaurand et al. 1999, 2004; Saito et al. 2012; Shariatgorji et al. 2014).

13.4 Secondary Ion Mass Spectrometry (SIMS) Imaging

Among all the MSI techniques, SIMS is the technique that was described the earliest, in 1910 and in 1960s, and has been used to image tissues for over three decades (Fletcher et al. 2011; Levi-Setti et al. 1985). SIMS imaging is often used to study

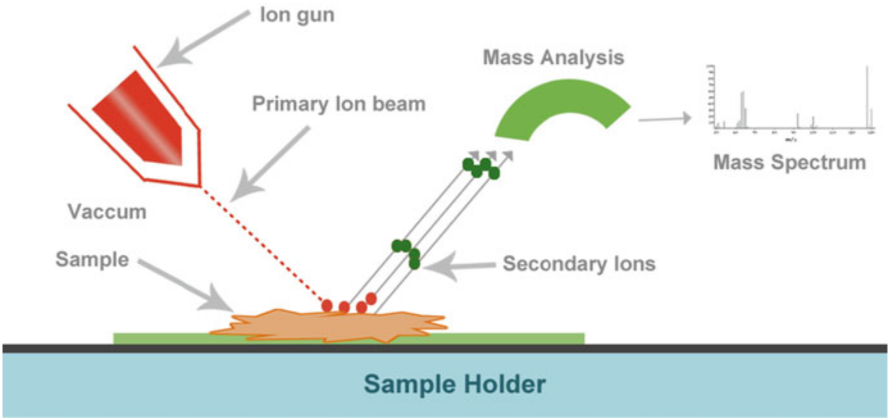


Fig. 13.2 Schematic representation of SIMS MSI process

the distribution of atoms and small molecules in tissues and even single cells at spatial resolutions below a micron.

SIMS is based on the principle that when a sample is bombarded with an electrostatically focused primary ion beam that can penetrate several nanometers into the sample, it can cause the ejection, or sputtering, of secondary species (electrons, neutrals, and ions) from the sample (Fletcher and Vickerman 2010; Vickerman 2011). The secondary ions can be electrostatically collected and mass analysed (Fig. 13.2). The penetration depth and amount of ejected matter is determined largely by selection of the specific primary ion source (Szakal et al. 2006). SIMS technique provides the advantages of measuring native and exogenous compounds in biological samples on a sub-cellular scale as the primary ion beam can be focussed onto a single cell. Two different approaches have been described for SIMS technique, static mode and dynamic modes of measurement. The dynamic mode uses high ion doses to quickly erode sample surfaces and obliterate molecular species making it best suited for elemental and isotopic studies, and depth profiling. The static mode uses low ion doses (less than 1 % of the surface molecules are impacted by a primary ion) to generate larger mass fragments and thus allows the study of all of the elements as well as molecular and molecular fragment ions. Both static and dynamic SIMS has been used successfully to image the distribution of chemicals across individual cells (Chilkoti et al. 1993; Fletcher and Vickerman 2010; Vickerman 2011).

Just like in MALDI-MSI, the first step in a SIMS measurement is generation of ions. Since the choice of the primary ion beam determines the overall performance, selection of appropriate source is very critical. A wide variety of primary ion sources have been described till date for a variety of samples, and of these the most common sources used for biological samples are Ga^+, Au_x^+, Bi_x^+ and C_{60}^+. Besides primary ion sources, the size of the beam and the primary ion flux, which are typically interdependent, influence the spatial resolution and duration of an imaging experiment.

The primary ion beam is readily focused on well below a micrometer in diameter for most sources, and is rastered across the sample under vacuum. The ejected secondary ions mass spectra is collected and analysed to get the spatial resolution (Chandra and Morrison 1992; Davies and Lynn 1990). Because of this technology, SIMS offers two major advantages over MALDI MSI, the first being the resolution that is achieved: SIMS can produce pixels on the order of 300 nm or so, compared to, at best, 1 mm with MALDI. The other is molecular depth profiling, where one can "dig into" the primary ion collision-induced craters in a sample and map its molecular composition in three dimensions. Some of the other advantages of using SIMS technique are that it requires very less sample or even samples with a low concentration levels can also be analysed. Also, unlike MALDI, SIMS does not require tedious sample preparation steps. There are few limitations in SIMS technique too: firstly, the material sputtered from the sample surface consists not only of mono-atomic ions but molecular species that in places can dominate the mass spectrum, making analysis of some elements impossible. Also, the sputtering process itself is poorly understood and there is no method to predict the secondary ionisation process. Thus a suitable standard that is close to the composition of the sample has to be used to obtain quantitative information. Another major limitation of SIMS is that the sample must be compatible with ultra high vacuum conditions (Davies and Lynn 1990; Fletcher et al. 2008; Ye et al. 2011).

13.5 Desorption Electro Spray Ionization (DESI) Imaging

DESI technique is a more recent development applied to imaging where samples are examined in the ambient environment with minimal pre-treatment. DESI was first developed in the laboratory of Professor R. Graham Cooks at Purdue University and later commercialised by Prosolia, Inc., (Takats et al. 2004). DESI was developed as alternate methods for the direct analysis of tissues at atmospheric pressure. The method requires minimal adaptation to existing mass spectrometers and unlike the previously discussed ionization methods, the DESI ionization source is relatively simple and may be readily constructed in house (Ye et al. 2011).

In a typical DESI-MS imaging experiment, a spray of electrically charged particles is directed towards the sample a few millimetres away with a help of voltage applied to the sample holder. When the spray impacts the sample, a thin layer of solvent is formed into which the analytes may dissolve. As other primary droplets arrive at the sample surface, they splash secondary micro-droplets containing the dissolved analytes from the solvent film and this process is termed as "droplet pickup". This is followed by the standard electrospray solvent evaporation processes, and finally the production of dry ions of analyte. These secondary ions travel through air under atmospheric pressure and is delivered to the mass spectrometer through a heated extended capillary system (Ye et al. 2011) (Fig. 13.3). In a regular DESI experimental set up the average velocity of the primary droplets is about 120 m/s, with an average diameter of about 3 μm and simulated DESI process shows

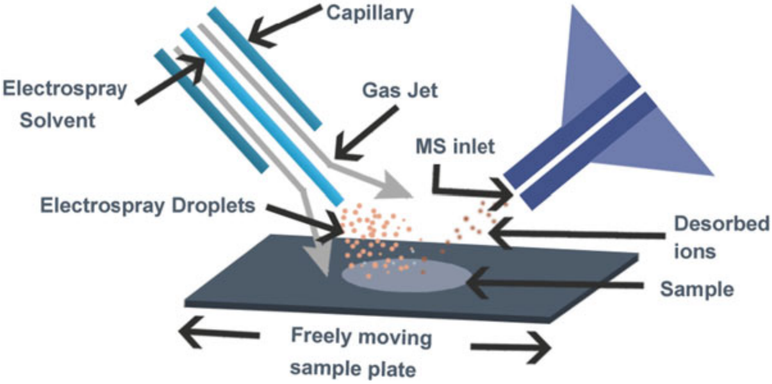


Fig. 13.3 Schematic representation of DESI MSI process

micro-droplets in the range of 0.8–3.3 μm, from a single droplet-thin film collision event. The most common samples for DESI MSI are biological tissue sections (fresh or frozen mounted on glass slide), tissue extracts and bacterial colonies (Dill et al. 2011; Eberlin et al. 2011; Kertesz and Van Berkel 2008; Miao and Chen 2009).

DESI MSI offers the advantages of direct analysis of biological samples, such as tissues, within seconds and with little or no sample preparation; it does not require the addition of a matrix and it is conducted outside the mass spectrometer at atmospheric pressure and it does not require vacuum. DESI also offers a rapid and high throughput results when compared to MALDI or SIMS MSI. The soft ionization method in DESI helps in detecting molecules that are intact and thus DESI MSI have been employed for in vivo imaging studies (Wiseman et al. 2008a). MALDI and DESI are two complementary methods in that, MALDI is primarily suited for detection of large molecules such as peptides and proteins, and DESI is well suited for detection of small molecules such as lipids, metabolites and drug molecules, however, MALDI and SIMS provides higher resolution than that of DESI (Takats et al. 2008; Wiseman et al. 2008a). DESI is also essentially label free that can be performed with basic instrumentation requirements (Takats et al. 2004). However, DESI is not free from limitations, some of the setbacks in DESI are: DESI cannot generally desorb molecules that are strongly bound to surfaces; there are low signals from molecules with low ionisation efficiency; and the spatial resolution is currently limited to approximately 100 μm (Ye et al. 2011).

13.6 Rapid Evaporative Ionization Mass Spectrometry (REIMS) Imaging

Traditional desorption/ionization techniques such as MALDI, SIMS, DESI are not suitable for evaluation of living tissue in-situ and in-vivo. Thermal evaporation helps in obtaining gaseous molecular ions from *in-situ* tissue using rapid thermal

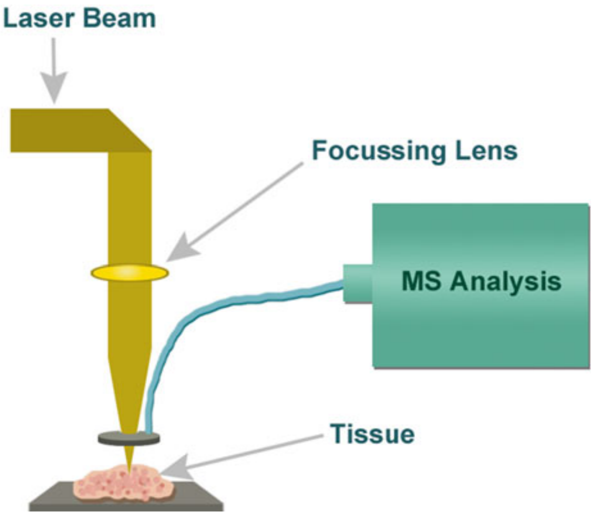


Fig. 13.4 Schematic representation of REIMS MSI process

evaporation. Various thermal evaporation techniques have been developed including thermo spray ionization. The rate of generation of gaseous molecular ions with thermal evaporation is compared to desorption techniques and predominantly contains phospholipids. These phospholipids and other molecular ions are evaluated using mass spectrometers and this technique is called Rapid evaporative ionization mass spectrometry (REIMS). Thermal evaporation techniques use high-frequency electric current (electro surgery) and laser for photo thermal effect (laser surgery). These surgical techniques are coupled with the REIMS principles to develop surgical knifes that result in thermal evaporation of tissue in-vivo. The surface of the tissue in contact with the surgical electrode or laser undergoes thermal evaporation. The resulting ions in gaseous form are transferred to a mass spectrometer using a venturi air jet pump (Fig. 13.4). The combination of REIMS with electro/laser surgery has been used for characterization of proteomic signature of tissues in-vivo, and this technique is called iKnife or Intelligent Knife. Results have shown strong correlation to distinct histological and histopathological tissue sub-types and also post-operative histological grading (Strittmatter et al. 2013).

13.7 Targets in MSI

13.7.1 MSI of Peptides and Proteins

The use of MS for the molecular analysis of elements and small organic molecules has been known for decades like the SIMS technology that dates back to more than 20 years. However, these techniques have not been effective for the image analysis of polypeptides and proteins. Of all the existing mass spectrometry imaging techniques, MALDI-MS imaging has proved to be the most useful for determining the localization of proteins and peptides (Chaurand et al. 1999, 2002; Strittmatter et al. 2013) directly in tissue samples for mid- to low molecular weight proteomes. Because this technology analyzes intact tissue, avoiding homogenization and separation steps, the spatial distribution of molecules within the tissue is preserved. Other methods like SIMS have been reported for analysis of peptides but it has only limited applications in the imaging of proteins and peptides because its mechanism of ionization causes extensive fragmentation of the target compounds, which makes their identification difficult. However, the use of a softer ionization method based on a Bi-cluster ion source makes it possible to analyze peptides with masses of up to 2 kDa by TOF-SIMS imaging (Chandra and Morrison 1992; Rabbani et al. 2011; Shariatgorji et al. 2014; Ye et al. 2011). The advantages of MALDI MSI like softer ionisation, direct tissue analysis etc., have made it the most popular MS technique for the imaging of proteins (Caprioli et al. 1997; Chaurand et al. 2002).

The first MALDI MSI for the analysis of peptides and proteins was reported in 1997 and since then has been applied to a wide variety of different tissues and analytical and clinical problems (Ye et al. 2011). A major focus of protein imaging has been in the area of cancer where studies have been carried out to improve molecular classification of grade, help predict clinical outcome, and examine molecular tumor margins. MALDI MSI has been recently applied to whole animal sections to examine protein, drug, and metabolite distributions (Goodwin et al. 2008). It has also been used for the 3D reconstruction of protein and peptide images within brain structures that can be correlated with standard MRI technique (Thiele et al. 2014). The use of MALDI MS technique to analyse banked human tissue samples that are formalin-fixed and paraffin-embedded (FFPE) has significant value for clinical diagnostics as often archives of FFPE tissues are associated with detailed patient information and can thus offer great potential for large scale studies on disease markers (Seeley and Caprioli 2008).

13.7.2 MSI of Lipids

Lipids have relatively high abundance in biological samples, and they are involved in numerous cellular processes and disease processes. Lipids were one of the first classes of compounds to be examined in MS imaging studies and MSI was used to identify the link of lipids and membranes to proteins and peptides (Shariatgorji et al. 2014).

MSI strategies like SIMS, MALDI, and DESI have all been reported to map the localization of lipids. Of these, SIMS has proven to be capable of mapping the localization of diverse lipids, including glycerophospholipids, sterol lipids, and sphingolipids with sub-micron spatial resolution (Borner et al. 2006; Nygren et al. 2005; Sjovall et al. 2004). However, its utility is limited by its tendency to cause extensive fragmentation of the target analytes and its comparatively low sensitivity in terms of secondary ion yields. SIMS imaging has been reported to investigate lipid distributions within mouse brain sections where cholesterol, sulfatides, and phosphatidylcholines were all identified (Sjovall et al. 2004). Because of the remarkable spatial resolution of SIMS, it is one of the few MSI techniques that can be used to visualize and identify individual lipids at the cellular and subcellular levels in cells such as neurons (Yang et al. 2010).

DESI MSI has also been reported for imaging involved in the visualization of lipids in rat brain tissue sections (Ifa et al. 2008). Depending on the solvents used and the nature of the substrate, spatial resolutions of around 200 mm could be achieved. However, it has been reported that imaging with a spatial resolution of around 12 mm is possible with a nano- DESI instrument (Laskin et al. (2012)). DESI-MSI has also been used to analyze and characterize the lipid profiles of different human astrocytoma subtypes, showing that some marker lipids have different abundances in different subtypes (Eberlin et al. 2010).

MALDI-MSI has been used extensively to map the distribution of lipids in a wide range of organs, including the brain. Benabdellah et al. compared MALDI and SIMS imaging for the visualization of rat brain lipids. It was concluded that MALDI-MS imaging is a robust and reproducible technique provided that care is taken during sample preparation and matrix application. The two techniques were found to have different advantages: MALDI-MS imaging was capable of nM sensitivity, whereas SIMS was able to achieve sub-mm spatial resolutions (Benabdellah et al. 2010).

13.7.3 MSI of Pharmaceuticals

The imaging of pharmaceuticals and their metabolites in their target sites is a very important source of information in drug discovery and development that can provide information on pharmacokinetics, toxicology, and ADME (absorption, distribution, metabolism, excretion). MSI strategies like MALDI, SIMS, and DESI techniques have all been used in mapping the distribution of pharmaceuticals. However, because of their greater sensitivity and softer ionization mechanisms, MALDI and DESI techniques are better suited than SIMS for the MSI of drugs and their metabolites in their native states without the need for labelling and several studies have reported their usage to image drug molecules and metabolite distributions in tissue sections (Shariatgorji et al. 2014).

Hsieh et al, have compared MALDI MSI with autoradiography and have shown consistent results in both methods when studying the rat brain tissues for clozapine drugs distribution (Hsieh et al. 2006). MALDI-MS imaging has also been shown to be able to directly and quantitatively determine the absolute concentrations of target compounds in specific regions of tissue sections (Goodwin et al. 2011; Nilsson et al. 2010). More recently MALDI MSI has also been used as a powerful tool in PET ligand research and development (Goodwin et al. 2011). MALDI-MS/MS imaging has also been used for visualizing the spatial distribution of astemizole and its primary metabolites in rat brain tissues to study the side effects of the drug in the central nervous system (Hsieh et al. 2010). More recently MALDI has been reported to visualize drug penetration across the blood brain barrier without molecular labelling and has been validated against the standard fluorescent microscopic examinations (Liu et al. 2013).

The ambient ionization method DESI-MSI is another technique used for imaging drugs and their metabolites. DESI has been used to perform direct, high-throughput imaging of clozapine drug to directly determine its distribution within histological brain sections. In this study, DESI was used for quantitative detection of the abundance of clozapine in brain samples (and also lung, kidney, and testis samples) and were compared with those obtained by conventional LC-MS/MS methods. The two data sets were shown to be in good agreement, which suggests that DESI imaging can be useful for the direct and quantitative detection of drugs and drug metabolites in biological tissue samples (Liu et al. 2013; Wiseman et al. 2008b).

13.7.4 MSI of Neurotransmitters and Endogenous Metabolites

Neurotransmitters are endogenous chemicals that transmit signals from a neuron to a target cell across a synapse. The spatial localization and molecular distribution of neurotransmitters as well as endogenous metabolites within biological organisms is of tremendous interest to neuroscientists. Application of MSI for studying distribution of neurotransmitters metabolites can be challenging because of the smaller molecular sizes, lesser abundance and faster molecular turnover of the targets as well as the samples studied are usually brain tissue sections that are rich in lipids.

There are a number of significant technical difficulties associated with MALDI-MS imaging of neurotransmitters, and considerable effort has been invested into finding ways of addressing them. A modification of MALDI using surface-assisted laser desorption ionization in place of a chemical matrix proved to be an alternative way of overcoming the matrix peak interference problem in MALDI-MS (Shariatgorji et al. 2009). Titanium dioxide nanoparticles have been used in a study that mapped the distribution of endogenous low molecular weight gamma amino butyric acid in mouse brain tissue sections (Shrivas et al. 2011).

In 2004, Touboul et al., developed a SIMS technique with a bismuth-based cluster ion source to visualize the distribution of lipids and some abundant small molecules (including cholesterol) in rat brain sections (Touboul et al. 2004).

Catecholamines (including epinephrine and norepinephrine) have been imaged by DESI-MS in porcine and rabbit adrenal glands. DESI-MS has also been successfully used to image cholesterol in adrenal gland and mouse brain samples (Wu et al. 2009, 2010).

MALDI MSI approach has been shown to be capable of simultaneous visualization of adenosine nucleotides in order to provide information on energy production and consumption in brain tissue sections. The method proved to be capable of directly detecting and identifying 13 primary metabolites in rat brain sections, at a spatial resolution of 50 mm (Benabdellah et al. 2009). More recently, high-resolution and high-accuracy mass spectrometers (HRMS) have been developed for the MSI of metabolites and neurotransmitters in rodent and crustacean central nervous systems (Ye et al. 2013). These open up a new field for studying neurotransmitters and endogenous metabolites *in situ* using MSI techniques.

13.7.5 MSI of Inorganic Ions

Metal ions have important roles in many signalling and metabolic pathways due to their diverse redox properties and varied coordination chemistry. Potassium, calcium, and sodium have functions in signal transduction, synaptic transmission, plasticity and cell excitability. Transition metals such as zinc, copper, manganese, and iron have essential roles in neurotransmitter synthesis, the regulation of synaptic transmission, and brain development (Shariatgorji et al. 2014).

SIMS technique has a long history as a highly sensitive technique for the imaging of inorganic ions in biological tissues with a high spatial resolution. It has been used to image sodium, potassium, magnesium, and calcium ions in retinal tissues (Kim et al. 2008) and aluminium in human brain sections (Candy et al. 1992). More recently MALDI-MSI was used to study the distribution of potassium ions in sagittal sections of rat brains at a spatial resolution of 100 mm (Shariatgorji et al. 2014).

13.8 MSI Data Analysis Software

13.8.1 Open MSI

Different Mass Spectroscopy Imaging (MSI) techniques discussed here generate enormous volumes of data and often the data is produced from different mass spectroscopy platforms. This makes analysis, interpretation of MSI data a considerably complex and resource intensive. MSI imaging typically produces spatial data over the tissue matrix containing the profile of intensity values over a range of mass to charge (m/z) values. The high spatial precision generates large volumes of highly complex data with each experiment (Rompp et al. 2011). Human Proteomics Organization (HUPO) has launched the Proteomics Standards Initiative (PSI) that supports development of open standards for data storage for MSI data such as imzML (Schramm et al. 2012; Shariatgorji et al. 2014), mzML (Martens et al. 2011) etc.

OpenMSI is a software platform developed to address these issues with MSI data and offers an extensible high density storage format which allows object oriented access via a simple yet extensively granular access to data via web API (Rubel et al. 2013). OpenMSI framework is conceptualized and maintained by Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

OpenMSI data format is built according to HDF5 data model. HDF (Hierarchical Data Format) allows storing complex data structures along with proprietary data items along with metadata. The root entry in openMSI HDF5 data model is/. Each mass spectroscopy imaging experiment is stored as/entry_# group. Metadata related to the experiment, instrument etc., are also stored within the/entry_# group. Raw MSI data and derived analysis results are stored within/ entry_#/data_# group. OpenMSI uses hybrid chunking to allow rapid selective data operations allowing mining of large and complex MSI data sets. High compression ratios are achieved using gzip compression in openMSI data sets. To improve the speed of orthogonal access to image and spectral data extensive data replication is used.

The OpenMSI Web API allows granular access to the underlying spectral and image data. It consists of five functions namely, qmetadata, qmz, qslice, qspectrum, and qcube. These functions can be used to get full access to the metadata, spectra, image data and raw MSI and derived analysis data. These methods are encoded in URL patterns that can be used to query the data repository with very fast turnaround times.

All data for openMSI framework is hosted on servers at National Energy Research Scientific Computing Center (NERSC). The NERSC resource provides high compute power and handles all data operations and acts as a data repository for openMSI data.

13.9 Conclusion and Future Perspectives

Many great advances have been made in the field of MSI to resolve molecular species in various types of biological samples. MSI offers technique for integrating biomolecular information that also provides chemical insights into the biomolecular associations between different groups of chemicals in tissues and shows how these associations can be affected by disease or by administration of a drug. On the other hand, MSI also presents significant hurdles like data sets generated are extremely complex as they consist of hundreds or thousands of molecule-specific images and figuring out how to work with those data, and especially making sense of them, is particularly challenging.

The power and capabilities of MS imaging technologies have increased substantially over the last decade, but further improvements are still needed, particularly with respect to sensitivity, quantitation, spatial resolution, the analysis of high molecular weight substances, seamless integration in multimodal experiments, compound identification, and throughput. Such improvements can be confidently expected because the technique is still young and is undergoing rapid development.

Acknowledgements We would like to thank Nissi Anthony and Siva Raman PN for helping in manuscript preparation.

References

Aebersold R, Goodlett DR. Mass spectrometry in proteomics. Chem Rev. 2001;101:269–95.

Benabdellah F, Touboul D, Brunelle A, Laprevote O. In situ primary metabolites localization on a rat brain section by chemical mass spectrometry imaging. Anal Chem. 2009;81:5557–60.

Benabdellah F, Seyer A, Quinton L, Touboul D, Brunelle A, Laprevote O. Mass spectrometry imaging of rat brain sections: nanomolar sensitivity with MALDI versus nanometer resolution by TOF-SIMS. Anal Bioanal Chem. 2010;396:151–62.

Borner K, Nygren H, Hagenhoff B, Malmberg P, Tallarek E, Mansson JE. Distribution of cholesterol and galactosylceramide in rat cerebellar white matter. Biochim Biophys Acta. 2006;1761:335–44.

Candy JM, Oakley AE, Mountfort SA, Taylor GA, Morris CM, Bishop HE, Edwardson JA. The imaging and quantification of aluminium in the human brain using dynamic secondary ion mass spectrometry (SIMS). Biol Cell. 1992;74:109–18.

Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. Anal Chem. 1997;69:4751–60.

Chandra S, Morrison GH. Sample preparation of animal tissues and cell cultures for secondary ion mass spectrometry (SIMS) microscopy. Biol Cell. 1992;74:31–42.

Chaurand P, Stoeckli M, Caprioli RM. Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. Anal Chem. 1999;71:5263–70.

Chaurand P, Schwartz SA, Caprioli RM. Imaging mass spectrometry: a new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections. Curr Opin Chem Biol. 2002;6:676–81.

Chaurand P, Schwartz SA, Capriolo RM. Profiling and imaging proteins in tissue sections by MS. Anal Chem. 2004;76:87A–93.

Chaurand P, Schwartz SA, Reyzer ML, Caprioli RM. Imaging mass spectrometry: principles and potentials. Toxicol Pathol. 2005;33:92–101.

Chilkoti A, Ratner BD, Briggs D. Static secondary ion mass spectrometric investigation of the surface chemistry of organic plasma-deposited films created from oxygen-containing precursors. 3. Multivariate statistical modeling. Anal Chem. 1993;65:1736–45.

Davies MC, Lynn RA. A review: secondary ion mass spectrometry (SIMS) of polymeric biomaterials. Clin Mater. 1990;5:97–114.

Dill AL, Eberlin LS, Ifa DR, Cooks RG. Perspectives in imaging using mass spectrometry. Chem Commun (Camb). 2011;47:2741–6.

Eberlin LS, Dill AL, Golby AJ, Ligon KL, Wiseman JM, Cooks RG, Agar NY. Discrimination of human astrocytoma subtypes by lipid analysis using desorption electrospray ionization imaging mass spectrometry. Angew Chem Int Ed Engl. 2010;49:5953–6.

Eberlin LS, Liu X, Ferreira CR, Santagata S, Agar NY, Cooks RG. Desorption electrospray ionization then MALDI mass spectrometry imaging of lipid and protein distributions in single tissue sections. Anal Chem. 2011;83:8366–71.

Fletcher JS, Vickerman JC. A new SIMS paradigm for 2D and 3D molecular imaging of biosystems. Anal Bioanal Chem. 2010;396:85–104.

Fletcher JS, Rabbani S, Henderson A, Blenkinsopp P, Thompson SP, Lockyer NP, Vickerman JC. A new dynamic in mass spectral imaging of single biological cells. Anal Chem. 2008;80:9058–64.

Fletcher JS, Vickerman JC, Winograd N. Label free biochemical 2D and 3D imaging using secondary ion mass spectrometry. Curr Opin Chem Biol. 2011;15:733–40.

Fournier I, Day R, Salzet M. Direct analysis of neuropeptides by in situ MALDI-TOF mass spectrometry in the rat brain. Neuro Endocrinol Lett. 2003;24:9–14.

Goodwin RJ, Pennington SR, Pitt AR. Protein and peptides in pictures: imaging with MALDI mass spectrometry. Proteomics. 2008;8:3785–800.

Goodwin RJ, Mackay CL, Nilsson A, Harrison DJ, Farde L, Andren PE, Iverson SL. Qualitative and quantitative MALDI imaging of the positron emission tomography ligands raclopride (a D2 dopamine antagonist) and SCH 23390 (a D1 dopamine antagonist) in rat brain tissue sections using a solvent-free dry matrix application method. Anal Chem. 2011;83:9694–701.

Hillenkamp F, Unsold E, Kaufmann R, Nitsche R. Laser microprobe mass analysis of organic materials. Nature. 1975;256:119–20.

Hsieh Y, Casale R, Fukuda E, Chen J, Knemeyer I, Wingate J, Morrison R, Korfmacher W. Matrix-assisted laser desorption/ionization imaging mass spectrometry for direct measurement of clozapine in rat brain tissue. Rapid Commun Mass Spectrom. 2006;20:965–72.

Hsieh Y, Li F, Korfmacher WA. Mapping pharmaceuticals in rat brain sections using MALDI imaging mass spectrometry. Methods Mol Biol. 2010;656:147–58.

Ifa DR, Manicke NE, Rusine AL, Cooks RG. Quantitative analysis of small molecules by desorption electrospray ionization mass spectrometry from polytetrafluoroethylene surfaces. Rapid Commun Mass Spectrom. 2008;22:503–10.

Kertesz V, Van Berkel GJ. Improved desorption electrospray ionization mass spectrometry performance using edge sampling and a rotational sample stage. Rapid Commun Mass Spectrom. 2008;22:3846–50.

Kim JH, Kim JH, Ahn BJ, Park JH, Shon HK, Yu YS, Moon DW, Lee TG, Kim KW. Label-free calcium imaging in ischemic retinal tissue by TOF-SIMS. Biophys J. 2008;94:4095–102.

Laskin J, Heath BS, Roach PJ, Cazares L, Semmes OJ. Tissue imaging using nanospray desorption electrospray ionization mass spectrometry. Anal Chem. 2012;84:141–8.

Levi-Setti R, Crow G, Wang YL. Progress in high resolution scanning ion microscopy and secondary ion mass spectrometry imaging microanalysis. Scan Electron Microsc. 1985;(Pt 2):535-52.

Liu X, Ide JL, Norton I, Marchionni MA, Ebling MC, Wang LY, Davis E, Sauvageot CM, Kesari S, Kellersberger KA, Easterling ML, Santagata S, Stuart DD, Alberta J, Agar JN, Stiles CD, Agar NY. Molecular imaging of drug transit through the blood-brain barrier with MALDI mass spectrometry imaging. Sci Rep. 2013;3:2859.

Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz PA, Deutsch EW. mzML – a community standard for mass spectrometry data. Mol Cell Proteomics. 2011;10:R110.

Miao Z, Chen H. Direct analysis of liquid samples by desorption electrospray ionization-mass spectrometry (DESI-MS). J Am Soc Mass Spectrom. 2009;20:10–9.

Nilsson A, Fehniger TE, Gustavsson L, Andersson M, Kenne K, Marko-Varga G, Andren PE. Fine mapping the spatial distribution and concentration of unlabeled drugs within tissue micro-compartments using imaging mass spectrometry. PLoS One. 2010;5:e11411.

Nygren H, Borner K, Hagenhoff B, Malmberg P, Mansson JE. Localization of cholesterol, phosphocholine and galactosylceramide in rat cerebellar cortex with imaging TOF-SIMS equipped with a bismuth cluster ion source. Biochim Biophys Acta. 2005;1737:102–10.

Rabbani S, Barber AM, Fletcher JS, Lockyer NP, Vickerman JC. TOF-SIMS with argon gas cluster ion beams: a comparison with C60+. Anal Chem. 2011;83:3793–800.

Rompp A, Guenther S, Takats Z, Spengler B. Mass spectrometry imaging with high resolution in mass and space (HR(2) MSI) for reliable investigation of drug compound distributions on the cellular level. Anal Bioanal Chem. 2011;401:65–73.

Rubel O, Greiner A, Cholia S, Louie K, Bethel EW, Northen TR, Bowen BP. OpenMSI: a high-performance web-based platform for mass spectrometry imaging. Anal Chem. 2013;85:10354–61.

Saito Y, Waki M, Hameed S, Hayasaka T, Setou M. Development of imaging mass spectrometry. Biol Pharm Bull. 2012;35:1417–24.

Schramm T, Hester A, Klinkert I, Both JP, Heeren RM, Brunelle A, Laprevote O, Desbenoit N, Robbe MF, Stoeckli M, Spengler B, Rompp A. imzML – a common data format for the flexible exchange and processing of mass spectrometry imaging data. J Proteomics. 2012;75:5106–10.

Seeley EH, Caprioli RM. Molecular imaging of proteins in tissues by mass spectrometry. Proc Natl Acad Sci U S A. 2008;105:18126–31.

Seeley EH, Oppenheimer SR, Mi D, Chaurand P, Caprioli RM. Enhancement of protein sensitivity for MALDI imaging mass spectrometry after chemical treatment of tissue sections. J Am Soc Mass Spectrom. 2008;19:1069–77.

Shariatgorji M, Spacil Z, Maddalo G, Cardenas LB, Ilag LL. Matrix-free thin-layer chromatography/laser desorption ionization mass spectrometry for facile separation and identification of medicinal alkaloids. Rapid Commun Mass Spectrom. 2009;23:3655–60.

Shariatgorji M, Svenningsson P, Andren PE. Mass spectrometry imaging, an emerging technology in neuropsychopharmacology. Neuropsychopharmacology. 2014;39:34–49.

Shrivas K, Hayasaka T, Sugiura Y, Setou M. Method for simultaneous imaging of endogenous low molecular weight metabolites in mouse brain using TiO2 nanoparticles in nanoparticle-assisted laser desorption/ionization-imaging mass spectrometry. Anal Chem. 2011;83:7283–9.

Sjovall P, Lausmaa J, Johansson B. Mass spectrometric imaging of lipids in brain tissue. Anal Chem. 2004;76:4271–8.

Strittmatter N, Jones EA, Veselkov KA, Rebec M, Bundy JG, Takats Z. Analysis of intact bacteria using rapid evaporative ionisation mass spectrometry. Chem Commun (Camb). 2013;49:6188–90.

Szakal C, Kozole J, Russo Jr MF, Garrison BJ, Winograd N. Surface sensitivity in cluster-ion-induced sputtering. Phys Rev Lett. 2006;96:216104.

Takats Z, Wiseman JM, Gologan B, Cooks RG. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. Science. 2004;306:471–3.

Takats Z, Wiseman JM, Ifa DR, Cooks RG. Desorption electrospray ionization: proteomics studies by a method that bridges ESI and MALDI. CSH Protoc. 2008;2008:pdb.top37.

Thiele H, Heldmann S, Trede D, Strehlow J, Wirtz S, Dreher W, Berger J, Oetjen J, Kobarg JH, Fischer B, Maass P. 2D and 3D MALDI-imaging: conceptual strategies for visualization and data mining. Biochim Biophys Acta. 2014;1844:117–37.

Touboul D, Halgand F, Brunelle A, Kersting R, Tallarek E, Hagenhoff B, Laprevote O. Tissue molecular ion imaging by gold cluster ion bombardment. Anal Chem. 2004;76:1550–9.

Vickerman JC. Molecular imaging and depth profiling by mass spectrometry – SIMS, MALDI or DESI? Analyst. 2011;136:2199–217.

Wiseman JM, Ifa DR, Zhu Y, Kissinger CB, Manicke NE, Kissinger PT, Cooks RG. Desorption electrospray ionization mass spectrometry: imaging drugs and metabolites in tissues. Proc Natl Acad Sci U S A. 2008;105:18120–5.

Wu C, Ifa DR, Manicke NE, Cooks RG. Rapid, direct analysis of cholesterol by charge labeling in reactive desorption electrospray ionization. Anal Chem. 2009;81:7618–24.

Wu C, Ifa DR, Manicke NE, Cooks RG. Molecular imaging of adrenal gland by desorption electrospray ionization mass spectrometry. Analyst. 2010;135:28–32.

Yang HJ, Ishizaki I, Sanada N, Zaima N, Sugiura Y, Yao I, Ikegami K, Setou M. Detection of characteristic distributions of phospholipid head groups and fatty acids on neurite surface by time-of-flight secondary ion mass spectrometry. Med Mol Morphol. 2010;43:158–64.

Ye H, Greer T, Li L. From pixel to voxel: a deeper view of biological tissue by 3D mass spectral imaging. Bioanalysis. 2011;3:313–32.

Ye H, Wang J, Greer T, Strupat K, Li L. Visualizing neurotransmitters and metabolites in the central nervous system by high resolution and high accuracy mass spectrometric imaging. ACS Chem Neurosci. 2013;4:1049–56.

Index

G. Marko-Varga (ed.), *Genomics and Proteomics for Clinical Discovery and Development*, Translational Bioinformatics 6, DOI 10.1007/978-94-017-9202-8