Advances in Experimental Medicine and Biology 926

Ákos Végvári *Editor*

Proteogenomics

Springer

Advances in Experimental Medicine and Biology

Volume 926

More information about this series at http://www.springer.com/series/5584

Ákos Végvári
Editor

Proteogenomics

Springer

Editor
Ákos Végvári
Clinical Protein Science & Imaging,
Department of Medical
Bioengineering, Biomedical Center
Lund University
Lund, Sweden

Department of Pharmacology &
Toxicology
University of Texas Medical Branch
Galveston, TX, USA

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-3-319-42314-2 ISBN 978-3-319-42316-6 (eBook)
DOI 10.1007/978-3-319-42316-6

Library of Congress Control Number: 2016951213

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

The concept of proteogenomics, utilizing advances from the fields of proteomics and genomics, was introduced at around the time of the completion of the sequencing of the human genome. The emergence of proteogenomics is mainly due to the rapid development of two key technologies: high-throughput DNA sequencing and mass spectrometry-based proteomics. The ability to determine protein sequences by mass spectrometry has provided a unique tool to the identification and the verification of novel genes, predicted exons, and open reading frames. Consequently, proteogenomics has been used for genome annotation, including the validation of known or annotated protein-coding genes; the improvement of gene annotations assigning correct start sites; the mapping of signal peptides, proteolysis, and other posttranslational modifications (an important element of biological function that is not encoded directly in the genome); as well as the identification of splicing variants and mutant proteoforms often associated with disease progression.

Considering the rapid advancement in the field, it is perhaps appropriate to define proteogenomics as an intensive research area that investigates the correlations between proteomic data and their corresponding genomic and transcriptomic data, keeping the goal to improve our knowledge about life at the molecular level, which is a more complete view that has been initially suggested. The interplay between the two data streams of genomics and proteomics certainly allows for a better understanding of biological functions and molecular mechanisms in health and disease. Today, genome sequencing provides nearly complete coverage, including transcriptome profiling, while targeted proteomics can be focused on specific regions of the proteome and determine predicted proteins.

The goal of this book is to display this extended view on proteogenomics, depicting research areas where proteogenomics is actively playing an essential role and also highlighting some emerging research arenas without pretending to cover all fields of application. The chapters of this book offer the readers a general insight to the integrative analyses of various types of omics data and present advances within specific principles, such as next-generation sequencing of DNA, mRNA sequencing, ribosome profiling, as well as mass spectrometry- and antibody-based proteomics. The applications are selected to exemplify the great potential of proteogenomics to contribute to human disease research, particularly to cancer and personalized medicine.

Importantly, this book attempts to identify some common features that integrate the various fields and areas where intensive efforts should be made to drive research more efficiently in the near future. One of these is certainly bioinformatics, which has shown amazing power and development during the last couple of years and which is anticipated to provide powerful approaches to improve our ability to work with and combine the large data sets that genomics, transcriptomics, and proteomics generate.

At last, I would like to thank all the authors of this book for their exceptional contributions, sharing their expert views of the field, and presenting their original research. Their enthusiasm and timely delivery of their manuscripts helped me tremendously to realize this project. It is my sincere hope that the readers would enjoy this book as much as I enjoyed preparing it.

Galveston, TX, USA Ákos Végvári
March 1, 2016

Contents

1 Proteogenomic Tools and Approaches to Explore Protein Coding Landscapes of Eukaryotic Genomes

Dhirendra Kumar and Debasis Dash

Abstract

Proteogenomic strategies aim to refine genome-wide annotations of protein coding features by using actual protein level observations. Most of the currently applied proteogenomic approaches include integrative analysis of multiple types of high-throughput omics data, *e.g.*, genomics, transcriptomics, proteomics, etc. Recent efforts towards creating a human proteome map were primarily targeted to experimentally detect at least one protein product for each gene in the genome and extensively utilized proteogenomic approaches. The 14 year long wait to get a draft human proteome map, after completion of similar efforts to sequence the genome, explains the huge complexity and technical hurdles of such efforts. Further, the integrative analysis of large-scale multi-omics datasets inherent to these studies becomes a major bottleneck to their success. However, recent developments of various analysis tools and pipelines dedicated to proteogenomics reduce both the time and complexity of such analysis. Here, we summarize notable approaches, studies, software developments and their potential applications towards eukaryotic genome annotation and clinical proteogenomics.

Keywords

Shotgun proteomics • Peptide identification • RNA-Seq • HUPO • Genome annotation

1.1 Introduction

Biological systems are complex, self-replicable machineries of which major components are proteins. Understanding the dynamics of protein expression in these systems may lead to a better interpretation of the underlying mechanisms and

D. Kumar • D. Dash (✉)
G.N. Ramachandran Knowledge Centre for Genome Informatics, CSIR-Institute of Genomics and Integrative Biology, South Campus, Sukhdev Vihar, Mathura Road, Delhi 110025, India
e-mail: ddash@igib.res.in

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_1

the predictability of potential outcomes. However, the techniques for probing these proteome components are not completely unbiased, *i.e.*, knowledge of each component of the proteome is necessary and prerequisite to probe their expression. These proteomic techniques are largely dependent on mass spectrometry (MS) based shotgun proteomics. Mass spectra, containing mass to charge ratios and intensities for peptides and their fragments are searched against a database of known proteins to identify the expressed proteins and their quantities (Eng et al. 2011). One of the limitations of this method lies in the database itself, against which the spectral data generated in MS are searched. A protein missing from the database cannot be probed for its expression, despite being present in the sample (Frank et al. 2007). Thus, for comprehensive proteome profiling, the search database should be complete. However, most of these databases are neither complete nor error free (Kumar et al. 2016b). Proteogenomic techniques address this problem by designing custom databases to identify the errors and achieve the completeness of the proteome definition for any organism (Castellana and Bafna 2010; Nesvizhskii 2014). Contrary to the routine proteomic searches, proteogenomic databases include proteins beyond the annotated proteome. Proteins from any organism are generally annotated by computationally predicting protein coding genes in the genome. While largely correct, these predictions also contain several inaccuracies. Proteogenomics relies on the detection of unique peptides from the MS data to correct these inaccuracies and refine the protein annotations on a genome wide scale (Jaffe et al. 2004; Yates et al. 1995).

Although very useful, these approaches are full of conceptual and technical challenges (Castellana and Bafna 2010). The order of complexity of proteogenomic approaches varies for different organisms. For example, for a prokaryotic genome, a six frame translated genome database should represent almost all possible protein coding genomic regions (Armengaud 2013; Kelkar et al. 2011; Kumar et al. 2013, 2014, 2016a). However, in the case of complex eukaryotes it would represent only a fraction of the possible protein species arising from the genome (Tanner et al. 2007). This is primarily due to alternative splicing of transcripts and only a tiny fraction of the eukaryotic genome being protein coding. Alternatively, proteogenomic databases for eukaryotes, to discover novel protein isoforms, generally integrate high-throughput transcriptomic information to discover new proteins from MS data searches. The high error rate, a byproduct of searching an extremely large database, is one of the major concerns in most of these studies (Krug et al. 2013; Yadav et al. 2013). Another factor contributing to potential false positive identifications is genomic polymorphism between individual genomes and the reference genome. These individual polymorphisms may result in new peptides from known genes, which may be mapped incorrectly to other places in the genome, leading to incorrect assignment of novel translated genomic regions. Additionally, inferring the exact isoform expressed in a given biological state is a difficult task in eukaryotic proteogenomics. Since various proteogenomic studies utilize a translated transcriptome as search database, which comprises of sequences of several transcripts from the same gene, many of the peptide identifications are shared among multiple database entries. Inferring the expressed protein isoform/s from the identified peptide list then becomes a non-trivial exercise and if incorrect it may adversely affect the conclusions. In addition to these, proteogenomic approaches are compute resource intensive (Castellana and Bafna 2010). Modern day approaches integrate multiple layers of omics information to discover novel protein isoforms. Each of these omics datasets, for example genomics, transcriptomics, proteomics, etc., is difficult to analyze independently. Further, their integration requires multivariate analyses (Horvatovich et al. 2015; Zhang et al. 2014) and considerations of multiple possible explanations for the observation (Omenn et al. 2015).

The complexity of such an analysis is reflected in several of the recent studies. For example, even after a decade since the human genome got sequenced, the characterization of the human proteome was achieved only recently and only as a draft version (Kim et al. 2014; Wilhelm et al. 2014). Nearly 20 % of the defined human protein coding genes are yet to be characterized at the protein level. Several worldwide initiatives are underway to detect at least one protein product for each of the human protein coding genes (Deutsch et al. 2015; Kumar et al. 2015; Nilsson et al. 2015; Paik et al. 2015). Similar incomplete proteome scenario exists for other model organisms, like mouse (Brosch et al. 2011), rat (Kumar et al. 2016b; Low et al. 2013), zebrafish (Kelkar et al. 2014), corn (*Zea maize*) (Castellana et al. 2014), etc. Despite various advances in MS instrumentation and analysis methods, defining the protein coding fraction for any genome remains incomplete. While the dynamics of protein expression is certainly one of the causes, the limited sensitivity of the method to detect low abundant proteins remains an open challenge and a primary cause of not detecting many proteins. Complexity of data analysis is another bottleneck in the detection of many proteins. Proteogenomic analyses directly address this point but are yet to be adapted in mainstream proteomic practice. Several of the recent tools and software packages that have been developed for use in proteogenomic analyses should make it an easy to implement approach and should expand its applications. Here, we would describe various analysis tools and pipelines targeted for eukaryotic proteogenomic pipelines.

1.2 Basics of Proteogenomics

Proteomics allows probing the expression of proteins from biological samples in a high-throughput manner (Steen and Mann 2004). Peptides are identified from mass spectra by searching against a protein sequence database using a search engine (Geer et al. 2004; Yadav et al. 2011) and identified peptides are mapped back to protein sequences to infer the expressed proteins (Eng et al. 2011). Proteogenomic approaches integrate these large-scale peptide discoveries with genomics and transcriptomics data to refine or enrich the annotation of protein coding genes (Armengaud 2009). Novel peptides, identified from proteogenomics, may reveal translation at the intergenic, intronic or annotated untranslated regions (UTRs) which may facilitate discovery of new genes, exons, splice variants and mutated proteins. However, such an analysis would require creation of custom search databases which maximizes the representation of such novel proteoforms; isoforms of proteins. Figure 1.1 highlights various possible custom database approaches and associated potential discoveries. Recently, various software tools and pipelines have been developed which either create a custom database or provide an end to end solution for proteogenomic data analysis and conclusions. The most significant contribution of these software solutions is to expand the outreach of such approaches to a larger scientific community, in addition to reducing the technical complexity and potential errors.

1.3 Proteogenomics Software Tools and Pipelines

A typical proteogenomic analysis includes custom database creation, peptide identification, genomic mapping of identified peptides and inferring the corrected or new gene model. Several of the recently developed tools offer only a part of the proteogenomic analysis, whereas few pipelines offer a complete proteogenomic workflow imlementation. For example:

- **CustomProDB** (Wang and Zhang 2013), an R package that allows for the creation of custom proteogenomic databases by incorporating single nucleotide polymorphism information from a common variant call format (vcf) file or from RNA-Seq data
- **SpliceDB** (Burset et al. 2001) allows creation of highly sensitive yet compact splice graph database in FASTA format which can be search by any of the peptide identification tools
- **MSProGene** (Zickmann and Renard 2015) is another standalone application that allows creation of a sample specific search database from RNA-Seq data with network information of peptide sharing among the database entries

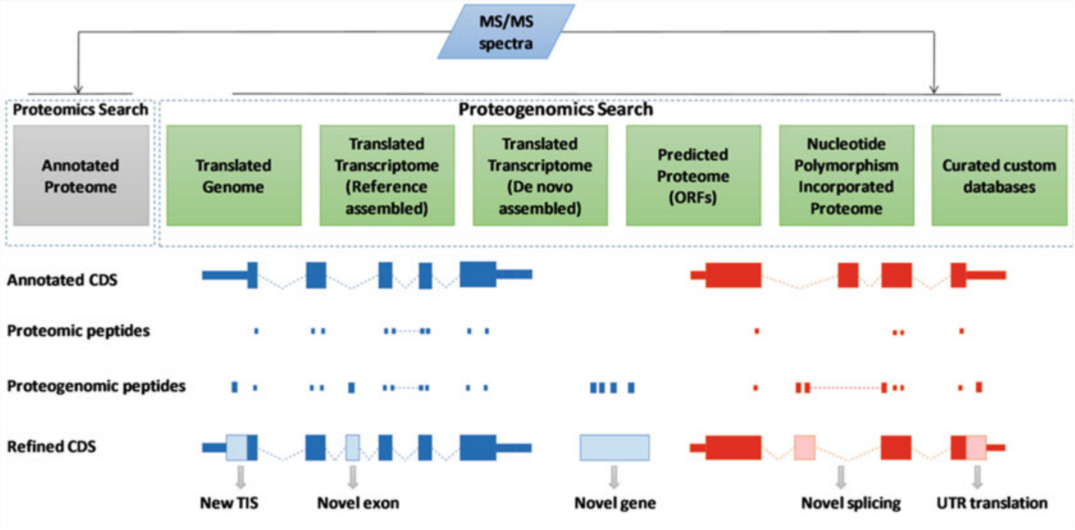


Fig. 1.1 Proteogenomic databases and refinement of genome annotations. *ORF* Open Reading Frame, *CDS* coding DNA sequences, *TIS* translation initiation site, *UTR* untranslated regions (Annotated). *Blue color rectangles* for CDS and peptides correspond to gene on positive strand whereas *red* colored ones for gene on negative strand

- **TheProteogenomic Mapping Tool** (Sanders et al. 2011) allows mapping of peptides back to the genome in a quick and effective manner
- **SpliceVista** (Zhu et al. 2014) is a Python package that maps identified peptides on all of the known splice-variants of proteins. It also allows integrated visualization of proteomics data with transcript information
- **dasHPPboard** (Tabas-Madrid et al. 2015) is a HUPO endorsed data integration platform which permits analysis and visualization of multiple omics datasets including proteomics
- **VESPA** (Peterson et al. 2012) is a JAVA based application that enables integrated visualization of transcriptomic and proteomics datasets in proteogenomic context
- **iPiG** (Kuhring and Renard 2012) allows integration of peptide identification into genome browser and thus, enables concurrent analysis of multiple omics information
- **PGx** (Askenazi et al. 2015), a recent tool converts peptide identifications into browser extensible format (BED) which contain genomic co-ordinates of features and can be visualized in genome browsers like UCSC
- Among the earliest proteogenomic pipelines, **Genome Annotating Proteomic Pipeline (GAPP)** (Shadforth et al. 2006) was designed specific to the human genome. This web based application improved the annotation for various genes by analyzing publicly available proteomics data. However, this pipeline is no longer active for use
- **PepLine** (Ferro et al. 2008) is standalone software for genome annotation which is independent of database search method. It rather relies on a hybrid tag based search to identify peptide tags and then maps and clusters these tags back to genome to discover potential translated regions. Due to the suspected low sensitivity and high-error rates of tag based peptide detection and genome mapping approach, it has only seen limited application in proteogenomics research
- **Peppy** (Risk et al. 2013) is one of the earliest developed pipelines for proteogenomic analysis. It is a fast and automated framework for quickly searching MS data against the extremely large eukaryotic genome translated databases to discover novel translated regions. Use of advanced computational methods in this tool makes proteogenomic searches implementable on simple desktop even for higher eukaryotic genomes which generally

necessitate higher memory and compute infrastructure. Additionally, it allows a blind modification search to account for novel post translation modifications which otherwise are very difficult to detect by regular proteomics searches. Despite these positive features, Peppy has limited eukaryotic analyses application as a large fraction of novel proteins in eukaryotes originate from alternate splicing of transcripts which cannot be represented in a genome translated search database as implemented in this pipeline

- **Enosi** (Castellana et al. 2014) proteogenomic pipeline is comprised of two functionalities. First, SpliceDB tool (Burset et al. 2001) is used to create a comprehensive yet compact database of splice junctions from RNA-Seq reads. This fasta formatted splice graph database is then searched with MS data using MS-GF+ search engine (Kim and Pevzner 2014) which is a sensitive tool to detect more peptides. To evaluate novel proteogenomic events including splice junctions, Enosi utilizes a probabilistic scoring which takes into account the number of spectra and peptides assigned to the locus, the quality of the assigned peptide spectral matches and the shared mapping of the peptide. The *eventProb* probabilistic score allows Enosi to rank and filter the proteogenomic findings according to their confidence. Further, the framework can utilize *ab initio* gene predictions and RNA-Seq information to estimate the boundaries of alternate gene models which accommodate the identified novel peptides. Additionally, Enosi pipeline is fully automated software and utilizes multi-threading to speed up the MS data searches
- **PGTools** (Nagaraj et al. 2015) is an end to end solution which seamlessly integrates multiple components of proteogenomic analysis. It is an open source software suite which offers fully automated searches along with the meta-analysis and visualization of novel findings. It allows searches against multiple custom databases, *e.g.*, databases containing translated entries from transcripts, non-coding genes, UTRs, six frame translated genome, splice junctions and somatic variations. By enabling searches against cancer specific variations from COSMIC database and fusion proteins, PGTools also allows human cancer specific proteogenomic studies. Further, its multiple search engine approach adds sensitivity to the overall peptide detection process. However, due to differences in peptide detection confidence inherent to variable database sizes, result integration from these different databases presents new challenges. Additionally, the approach lacks the strength of individual or tissue specific proteogenomic searches as that from RNA-Seq data
- **ProteoAnnotator** (Ghali et al. 2014) is a recent, open source and powerful pipeline for proteogenomic discoveries from MS datasets. It addresses one of the common problems of proteomics and proteogenomics research: file format standards. The entire pipeline supports and exports HUman Proteomics Organization (HUPO) – Proteomics Standards Initiative (PSI) supported file formats like MzIdentML. Proteoannotator also allows multiple database searches but primarily relies on gene predictions. Searching MS data against gene predictions is an excellent approach for a newly sequenced genome primarily due to increased sensitivity of peptide detection attributable to small search database compared to genomic or transcriptomic databases. The pipeline also introduces a "non-canonical gene model score" calculation which allows to assign confidence values to novel discoveries and thus automated assessment of quality of novel findings. In addition to these new features, it also presents an automated framework which integrates multiple peptide search engines and comprehensive statistical algorithm, FDRscore for result integration. Although it is very effective for proteogenomically annotating new genomes, individual or sample based database searches are difficult to implement in this framework
- **Integrated transcriptomic-proteomic pipeline (ITP)** (Kumar et al. 2016b) is a recently published pipeline and comprises two analysis modules, each for transcriptomics and pro-

teomics data. The transcriptomic analysis module uses Tuxedo suite of tools to align and assemble RNA-Seq reads into transcripts by utilizing the reference genome. Second module creates a translated transcriptome database from the assembled transcripts and then searches mass spectra against this database using multiple search engines. Although the pipeline lacks an entirely automated structure for public use, the approach has several advantages. For example, using a reference genome guided transcriptome assembly provides a definitive transcript model for the discovered novel peptides and thus, proper reannotation of exon boundaries and coding splice variants are possible. Similarly, quantities of transcript isoforms may indicate most probable protein coding isoform despite extensive peptide sharing among isoforms. It also allows creation of tissue or individual specific search databases specifically useful in clinical studies. In addition to these, multiple search engines and FDRscore (Jones et al. 2009; Kumar et al. 2013) based result integration within the second module **EuGenoSuite**, maximize both the sensitivity and specificity of peptide detection. Identified peptides are also exported into gene transfer format (GTF) which can be easily integrated into most of the genome browsers and thus enabling easy visualization of novel regions

– **PPLine** (Krasnov et al. 2015) is a Python language based automated proteogenomic pipeline which integrates proteomics with exome sequencing and transcriptome sequencing technologies. Its major focus is to discover variant novel peptides resulting from single nucleotide polymorphism (SNP), insertions-deletions in the genomic DNA and due to alternative splicing. It integrates several tools to accurately call SNPs from exome sequencing reads, align RNA-Seq reads, assemble transcripts including splice junction isoforms from reads and then allows proteomics data searches against variant peptide database. This comprehensive software enables sample/tissue specific database creation and thus facilitates clinical proteogenomic analysis

– **GALAXY-P** (Jagtap et al. 2014) is among the few web-based frameworks for proteogenomics. Despite its web based implementation, it allows extensive analysis for eukaryotic genomes with flexibilities at every step of analysis. It extends the Galaxy bioinformatics framework for proteomics data analysis and allows user to create custom integrative analysis workflows. Default workflows within Galaxy-P allow MS data format conversion, creation of proteogenomic databases from various web resources, two step database search and statistical assessment of identified peptides, sequence similarity searches of novel findings, evaluation of peptide-spectral matches by visualization and comprehensive genomic visualization of novel peptides. The Galaxy framework allows smooth integration of various genomics and transcriptomics data analysis and with the Galaxy-P development, integration of proteomics with other omics datasets becomes easy to implement. For example, Sheynkman et al. (2014) developed three analysis workflows which enable proteomics data searching within Galaxy-P framework against single amino acid polymorphism (SAP) and splice variant database developed from RNA-Seq data

– **QUILTS** (Zhang et al. 2009) is a software to create individual specific human proteogenomic search databases by integrating SAP variations, splice variants, gene fusions to canonical protein sequences. Individual specific genomic and transcriptomic variations have been attributed to different diseases primarily cancers and thus, it should allow clinical proteogenomic studies focused to detect disease specific variants. However, it is limited to human only and does not allow similar analysis for other model organisms, used to study human diseases.

With so many alternatives, one compelling question still remains: Which one is the best? Although, there have not been many studies which compare the various pipelines available for eukaryotic proteogenomics, our recent study suggests that many of these are actually complementary in their results (Kumar et al. 2016b). We concluded that due to differences in their search database compositions ITP, Enosi, ProteoAnnotator and Peppy bring complementary peptide detections. Although, there are many technical challenges to run multiple proteogenomic pipelines on a large scale proteomic dataset, the strategy would help achieve a comprehensive catalogue of novel translation events across genome.

1.4 Future Perspectives

Although these tools have reduced the technical complexity of proteogenomic searches, quality assessment of novel discoveries still remains a formidable challenge. Many studies indicate the necessity of manual inspection of identified peptide spectrum matches to ascertain true identifications (Omenn et al. 2015). However, it is not feasible to implement manual inspection on large scale studies. Tools like Enosi and ProteoAnnotator devised automated scoring systems to evaluate the novel identifications separately for their authenticity, but a comprehensive statistical framework dedicated to large scale proteogenomic studies is still needed. For example, both of the studies claiming to achieve a draft human proteome map have been heavily criticized for their high number of "low quality" identifications, adding up to false positives (Ezkurdia et al. 2014). There have been few approaches suggested to overcome these hurdles (Shanmugam and Nesvizhskii 2015; Zhang et al. 2015). However, these are yet to be implemented in automated pipelines. Other than statistical attributes, false positives may also arise due to incorrect genomic mapping of identified peptides. The genome of an individual can vary considerably from the reference genomes at various places, characterized by genomic variations like SNPs, insertions and deletions. If these are not taken into account, many of the peptide identifications may be incorrectly assigned to novel loci. Proteogenomic pipelines need to include this consideration while evaluating a novel translated region.

Integration of other omics readouts in proteogenomic frameworks could also be extremely beneficial. Particularly, ribosome bound RNAs (Ribosome profiling), rather than entire transcriptome, to create a custom search database that would allow for a better profiling of translated proteins and thus a better genome annotation. The recently developed PROTEOFORMER (Crappe et al. 2014) pipeline integrates ribosome profiling with MS based proteomics and proteogenomics analysis and could be extremely useful in eukaryotic genome annotations. However, a similar integration in other existing pipelines would expand the reach of such methods. These pipelines also need to include provisions for unsequenced genomes. Custom *de novo* assembled transcriptomes may provide templates for proteome profiling from MS data (Brinkman et al. 2015). Proteogenomic pipelines need to be extended to include genome independent database creation, to facilitate similar analysis for unsequenced or partially sequenced genomes.

Proteogenomic analyses hold promise for human disease related studies as well. Recent studies suggest the potential of proteogenomics in the discovering novel candidates in different cancers (Alfaro et al. 2014; Rivers et al. 2014; Woo et al. 2014; Zhang et al. 2014). However, most of the existing pipelines do not consider disease related genetic components. Extending these analysis frameworks would not only benefit new studies, they would also assist in revisiting previous datasets for proteogenomic reanalysis.

Acknowledgements Authors would like to thank CSIR-IGIB for compute infrastruture and project BSC0121 for publication charges.

References

Alfaro, J. A., Sinha, A., Kislinger, T., & Boutros, P. C. (2014). Onco-proteogenomics: Cancer proteomics joins forces with genomics. *Nature Methods, 11*(11), 1107–1113. Available from: PM:25357240.

Armengaud, J. (2009). A perfect genome annotation is within reach with the proteomics and genomics alliance. *Current Opinion in Microbiology, 12*(3), 292–300. Available from: PM:19410500.

Armengaud, J. (2013). Microbiology and proteomics, getting the best of both worlds! *Environmental Microbiology, 15*(1), 12–23. Available from: PM:22708953.

Askenazi, M., Ruggles, K. V., & Fenyo, D. (2015). PGx: Putting peptides to BED. *Journal of Proteome Research, 15*(3), 795–799. Available from: PM:26638927.

Brinkman, D. L., Jia, X., Potriquet, J., Kumar, D., Dash, D., Kvaskoff, D., & Mulvenna, J. (2015). Transcriptome and venom proteome of the box jellyfish Chironex fleckeri. *BMC Genomics, 16*, 407. Available from: PM:26014501.

Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., & Hubbard, T. (2011). Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Research, 21*(5), 756–767. Available from: PM:21460061.

Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2001). SpliceDB: Database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Research, 29*(1), 255–259. Available from: PM:11125105.

Castellana, N., & Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of Proteomics, 73*(11), 2124–2135. Available from: PM:20620248.

Castellana, N. E., Shen, Z., He, Y., Walley, J. W., Cassidy, C. J., Briggs, S. P., & Bafna, V. (2014). An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. *Molecular & Cellular Proteomics, 13*(1), 157–167. Available from: PM:24142994.

Crappe, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De, K. S., De, M. E., De, M. T., Van, C. W., Van, D. P., & Menschaert, G. (2014). PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research, 43*(5), e29. Available from: PM:25510491.

Deutsch, E. W., Sun, Z., Campbell, D., Kusebauch, U., Chu, C. S., Mendoza, L., Shteynberg, D., Omenn, G. S., & Moritz, R. L. (2015). State of the human proteome in 2014/2015 as viewed through PeptideAtlas: Enhancing accuracy and coverage through the AtlasProphet. *Journal of Proteome Research, 14*(9), 3461–3473. Available from: PM:26139527.

Eng, J. K., Searle, B. C., Clauser, K. R., & Tabb, D. L. (2011). A face in the crowd: Recognizing peptides through database search. *Molecular Cellular Proteomics, 10*(11), R111. Available from: PM:21876205.

Ezkurdia, I., Vazquez, J., Valencia, A., & Tress, M. (2014). Analyzing the first drafts of the human proteome. *Journal of Proteome Research, 13*(8), 3854–3855. Available from: PM:25014353.

Ferro, M., Tardif, M., Reguer, E., Cahuzac, R., Bruley, C., Vermat, T., Nugues, E., Vigouroux, M., Vandenbrouck, Y., Garin, J., & Viari, A. (2008). PepLine: A software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. *Journal of Proteome Research, 7*(5), 1873–1883. Available from: PM:18348511.

Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A., & Pevzner, P. A. (2007). De novo peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research, 6*(1), 114–123. Available from: PM:17203955.

Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., & Bryant, S. H. (2004). Open mass spectrometry search algorithm. *Journal of Proteome Research, 3*(5), 958–964. Available from: PM:15473683.

Ghali, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Wastling, J., & Jones, A. R. (2014). ProteoAnnotator – Open source proteogenomics annotation software supporting PSI standards. *Proteomics, 14*(23–24), 2731–2741. Available from: PM:25297486.

Horvatovich, P., Lundberg, E. K., Chen, Y. J., Sung, T. Y., He, F., Nice, E. C., Goode, R. J., Yu, S., Ranganathan, S., Baker, M. S., Domont, G. B., Velasquez, E., Li, D., Liu, S., Wang, Q., He, Q. Y., Menon, R., Guan, Y., Corrales, F. J., Segura, V., Casal, J. I., Pascual-Montano, A., Albar, J. P., Fuentes, M., Gonzalez-Gonzalez, M., Diez, P., Ibarrola, N., Degano, R. M., Mohammed, Y., Borchers, C. H., Urbani, A., Soggiu, A., Yamamoto, T., Salekdeh, G. H., Archakov, A., Ponomarenko, E., Lisitsa, A., Lichti, C. F., Mostovenko, E., Kroes, R. A., Rezeli, M., Vegvari, A., Fehniger, T. E., Bischoff, R., Vizcaino, J. A., Deutsch, E. W., Lane, L., Nilsson, C. L., Marko-Varga, G., Omenn, G. S., Jeong, S. K., Lim, J. S., Paik, Y. K., & Hancock, W. S. (2015). Quest for missing proteins: Update 2015 on chromosome-centric human proteome project. *Journal of Proteome Research, 14*(9), 3415–3431. Available from: PM:26076068.

Jaffe, J. D., Berg, H. C., & Church, G. M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics, 4*(1), 59–77. Available from: PM:14730672.

Jagtap, P. D., Johnson, J. E., Onsongo, G., Sadler, F. W., Murray, K., Wang, Y., Shenykman, G. M., Bandhakavi, S., Smith, L. M., & Griffin, T. J. (2014). Flexible and accessible workflows for improved proteogenomic analysis using the galaxy framework. *Journal of Proteome Research, 13*(12), 5898–5908. Available from: PM:25301683.

Jones, A. R., Siepen, J. A., Hubbard, S. J., & Paton, N. W. (2009). Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics, 9*(5), 1220–1229. Available from: PM:19253293.

Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Yadav, A. K., Shrivastava, P., Marimuthu, A., Anand, S., Sundaram, H., Kingsbury, R., Harsha, H. C., Nair, B., Prasad, T. S., Chauhan, D. S., Katoch, K., Katoch, V. M., Kumar, P., Chaerkady, R., Ramachandran, S., Dash, D., & Pandey, A. (2011). Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. *Molecular*

Cellular Proteomics, 10(12), M111. Available from: PM:21969609.

Kelkar, D. S., Provost, E., Chaerkady, R., Muthusamy, B., Manda, S. S., Subbannayya, T., Selvan, L. D., Wang, C. H., Datta, K. K., Woo, S., Dwivedi, S. B., Renuse, S., Getnet, D., Huang, T. C., Kim, M. S., Pinto, S. M., Mitchell, C. J., Madugundu, A. K., Kumar, P., Sharma, J., Advani, J., Dey, G., Balakrishnan, L., Syed, N., Nanjappa, V., Subbannayya, Y., Goel, R., Prasad, T. S., Bafna, V., Sirdeshmukh, R., Gowda, H., Wang, C., Leach, S. D., & Pandey, A. (2014). Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. *Molecular Cellular Proteomics, 13*(11), 3184–3198. Available from: PM:25060758.

Kim, S., & Pevzner, P. A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications, 5*, 5277. Available from: PM:25358478.

Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T. C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H., & Pandey, A. (2014). A draft map of the human proteome. *Nature, 509*(7502), 575–581. Available from: PM:24870542.

Krasnov, G. S., Dmitriev, A. A., Kudryavtseva, A. V., Shargunov, A. V., Karpov, D. S., Uroshlev, L. A., Melnikova, N. V., Blinov, V. M., Poverennaya, E. V., Archakov, A. I., Lisitsa, A. V., & Ponomarenko, E. A. (2015). PPLine: An automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics. *Journal of Proteome Research, 14*(9), 3729–3737. Available from: PM:26147802.

Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., & Macek, B. (2013). Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Molecular Cellular Proteomics, 12*(11), 3420–3430. Available from: PM:23908556.

Kuhring, M., & Renard, B. Y. (2012). iPiG: Integrating peptide spectrum matches into genome browser visualizations. *PLoS One, 7*(12), e50246. Available from: PM:23226516.

Kumar, D., Yadav, A. K., Kadimi, P. K., Nagaraj, S. H., Grimmond, S. M., & Dash, D. (2013). Proteogenomic analysis of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline. *Molecular Cellular Proteomics, 12*(11), 3388–3397. Available from: PM:23882027.

Kumar, D., Mondal, A. K., Yadav, A. K., & Dash, D. (2014). Discovery of rare protein-coding genes in model methylotroph Methylobacterium extorquens AM1. *Proteomics, 14*(23–24), 2790–2794. Available from: PM:25158906.

Kumar, D., Jain, A., & Dash, D. (2015). Probing the missing human proteome: A computational perspective. *Journal of Proteome Research, 14*(12), 4949–4958. Available from: PM:26407240.

Kumar, D., Mondal, A. K., Kutum, R., & Dash, D. (2016a). Proteogenomics of rare taxonomic phyla: A prospective treasure trove of protein coding genes 2. *Proteomics, 16*(2), 226–240. Available from: PM:26773550.

Kumar, D., Yadav, A. K., Jia, X., Mulvenna, J., & Dash, D. (2016b). Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. *Molecular Cellular Proteomics, 15*(1), 329–339. Available from: PM:26560066.

Low, T. Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hubner, N., van Breukelen, B., Mohammed, S., Cuppen, E., Heck, A. J., & Guryev, V. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Reports, 5*(5), 1469–1478. Available from: PM:24290761.

Nagaraj, S. H., Waddell, N., Madugundu, A. K., Wood, S., Jones, A., Mandyam, R. A., Nones, K., Pearson, J. V., & Grimmond, S. M. (2015). PGTools: A software suite for proteogenomic data analysis and visualization. *Journal of Proteome Research, 14*(5), 2255–2266. Available from: PM:25760677.

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods, 11*(11), 1114–1125. Available from: PM:25357241.

Nilsson, C. L., Mostovenko, E., Lichti, C. F., Ruggles, K., Fenyo, D., Rosenbloom, K. R., Hancock, W. S., Paik, Y. K., Omenn, G. S., LaBaer, J., Kroes, R. A., Uhlen, M., Hober, S., Vegvari, A., Andren, P. E., Sulman, E. P., Lang, F. F., Fuentes, M., Carlsohn, E., Emmett, M. R., Moskal, J. R., Berven, F. S., Fehniger, T. E., & Marko-Varga, G. (2015). Use of ENCODE resources to characterize novel proteoforms and missing proteins in the human proteome. *Journal of Proteome Research, 14*(2), 603–608. Available from: PM:25369122.

Omenn, G. S., Lane, L., Lundberg, E. K., Beavis, R. C., Nesvizhskii, A. I., & Deutsch, E. W. (2015). Metrics for the human proteome project 2015: Progress on the human proteome and guidelines for high-confidence protein identification. *Journal of Proteome Research, 14*(9), 3452–3460. Available from: PM:26155816.

Paik, Y. K., Omenn, G. S., Overall, C. M., Deutsch, E. W., & Hancock, W. S. (2015). Recent advances in the chromosome-centric human proteome project: Missing proteins in the spot light. *Journal of Proteome Research, 14*(9), 3409–3414. Available from: PM:26337862.

Peterson, E. S., McCue, L. A., Schrimpe-Rutledge, A. C., Jensen, J. L., Walker, H., Kobold, M. A., Webb, S. R., Payne, S. H., Ansong, C., Adkins, J. N., Cannon, W. R., & Webb-Robertson, B. J. (2012). VESPA: Software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data. *BMC Genomics, 13*, 131. Available from: PM:22480257.

Risk, B. A., Spitzer, W. J., & Giddings, M. C. (2013). Peppy: Proteogenomic search software. *Journal of Proteome Research, 12*(6), 3019–3025. Available from: PM:23614390.

Rivers, R. C., Kinsinger, C., Boja, E. S., Hiltke, T., Mesri, M., & Rodriguez, H. (2014). Linking cancer genome to proteome: NCI's investment into proteogenomics. *Proteomics, 14*(23–24), 2633–2636. Available from: PM:25187343.

Sanders, W. S., Wang, N., Bridges, S. M., Malone, B. M., Dandass, Y. S., McCarthy, F. M., Nanduri, B., Lawrence, M. L., & Burgess, S. C. (2011). The proteogenomic mapping tool. *BMC Bioinformatics, 12*, 115. Available from: PM:21513508.

Shadforth, I., Xu, W., Crowther, D., & Bessant, C. (2006). GAPP: A fully automated software for the confident identification of human peptides from tandem mass spectra. *Journal of Proteome Research, 5*(10), 2849–2852. Available from: PM:17022656.

Shanmugam, A. K., & Nesvizhskii, A. I. (2015). Effective leveraging of targeted search spaces for improving peptide identification in tandem mass spectrometry based proteomics. *Journal of Proteome Research, 14*(12), 5169–5178. Available from: PM:26569054.

Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., Griffin, T. J., & Smith, L. M. (2014). Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics, 15*, 703. Available from: PM:25149441.

Steen, H., & Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology, 5*(9), 699–711. Available from: PM:15340378.

Tabas-Madrid, D., Alves-Cruzeiro, J., Segura, V., Guruceaga, E., Vialas, V., Prieto, G., Garcia, C., Corrales, F. J., Albar, J. P., & Pascual-Montano, A. (2015). Proteogenomics dashboard for the human proteome project 1. *Journal of Proteome Research, 14*(9), 3738–3749. Available from: PM:26144527.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S. P., & Bafna, V. (2007). Improving gene annotation using peptide mass spectrometry. *Genome Research, 17*(2), 231–239. Available from: PM:17189379.

Wang, X., & Zhang, B. (2013). CustomProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search 1. *Bioinformatics, 29*(24), 3235–3237. Available from: PM:24058055.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas, G. A., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J. H., Bantscheff, M., Gerstmair, A., Faerber, F., & Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature, 509*(7502), 582–587. Available from: PM:24870543.

Woo, S., Cha, S. W., Na, S., Guest, C., Liu, T., Smith, R. D., Rodland, K. D., Payne, S., & Bafna, V. (2014). Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics, 14*(23–24), 2719–2730. Available from: PM:25263569.

Yadav, A. K., Kumar, D., & Dash, D. (2011). MassWiz: A novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. *Journal of Proteome Research, 10*(5), 2154–2160. Available from: PM:21417338.

Yadav, A. K., Kadimi, P. K., Kumar, D., & Dash, D. (2013). ProteoStats–a library for estimating false discovery rates in proteomics pipelines. *Bioinformatics, 29*(21), 2799–2800. Available from: PM:23962616.

Yates, J. R., III, Eng, J. K., & McCormack, A. L. (1995). Mining genomes: Correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Analytical Chemistry, 67*(18), 3202–3210. Available from: PM:8686885.

Zhang, G., Fenyo, D., & Neubert, T. A. (2009). Evaluation of the variation in sample preparation for comparative proteomics using stable isotope labeling by amino acids in cell culture. *Journal of Proteome Research, 8*(3), 1285–1292. Available from: PM:19140678.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J., & Liebler, D. C. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature, 513*(7518), 382–387. Available from: PM:25043054.

Zhang, K., Fu, Y., Zeng, W. F., He, K., Chi, H., Liu, C., Li, Y. C., Gao, Y., Xu, P., & He, S. M. (2015). A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics, 31*(20), 3249–3253. Available from: PM:26076724.

Zhu, Y., Hultin-Rosenberg, L., Forshed, J., Branca, R. M., Orre, L. M., & Lehtio, J. (2014). SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Molecular Cell Proteomics, 13*(6), 1552–1562. Available from: PM:24692640.

Zickmann, F., & Renard, B. Y. (2015). MSProGene: Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics, 31*(12), i106–i115. Available from: PM:26072472.

Next Generation Sequencing Data and Proteogenomics

2

Kelly V. Ruggles and David Fenyö

Abstract

The field of proteogenomics has been driven by combined advances in next-generation sequencing (NGS) and proteomic methods. NGS technologies are now both rapid and affordable, making it feasible to include sequencing in the clinic and academic research setting. Alongside the improvements in sequencing technologies, methods in high throughput proteomics have increased the depth of coverage and the speed of analysis. The integration of these data types using continuously evolving bioinformatics methods allows for improvements in gene and protein annotation, and a more comprehensive understanding of biological systems.

Keywords

Next generation sequencing • Proteogenomic integration • Bioinformatics • Peptide identification • Gene annotation

2.1 NGS Overview

NGS itself refers to a number of techniques, all of which perform massively parallel sequencing, in which millions of DNA fragments from a sample are sequenced at the same time (Muzzey et al. 2015). This produces a vast amount of data, in some cases adding up to 1 TB per run. With this level of data volume and faster data generation, bioinformatics has emerged as the true challenge in NGS data analysis and integration.

The most frequently used NGS methods at the DNA level are whole exome sequencing and whole genome sequencing (WGS). In whole

K.V. Ruggles
Department of Medicine, New York University
Medical Center, 550 First Avenue, New York,
NY 10016, USA

Center for Health Informatics and Bioinformatics,
New York University Medical Center, 227 East 30th
Street, New York, NY 10016, USA
e-mail: kelly.ruggles@nyumc.org

D. Fenyö (✉)
Institute for Systems Genetics, New York University
Medical Center, 430 East 29th Street, New York,
NY 10016, USA

Department of Biochemistry and Molecular
Pharmacology, New York University Medical Center,
550 First Avenue, New York, NY 10016, USA
e-mail: david@fenyolab.org

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_2

exome sequencing, only protein-coding regions of the genome are sequenced, removing the remaining ~99 % of the DNA and thereby significantly lowering the required time and cost. This method has been most often employed in studies of gene discovery and the identification of disease causing mutations. For WGS however, the entire genome is sequenced, which is useful for novel gene identification and for the analysis of non-coding regions including promoters and enhancers.

DNA NGS technologies have enabled researchers to detect differences between an experimental and a reference genome. These typically fall into two categories:

1. Large deletions/duplications (copy number variation (CNV))
2. Changes to the DNA sequence, also known as "variants", either as single nucleotide polymorphisms (SNPs) or short insertion/deletions (indels)

Both require alignment of NGS reads to a reference genome (Fig. 2.1).

NGS is also performed on RNA using RNA-Seq, a technique which is now frequently used in lieu of micro-arrays to assess gene expression. RNA-Seq enables researchers to investigate alternative splicing events, gene fusion events, SNPs and gene expression. The experimental procedure is similar to that of DNA sequencing, with an additional step of first deriving cDNA sequences from all RNA present in the sample.

Although different sequencing methods can produce different raw data types, these data are most often combined to create a FASTQ file, containing information on both sequence and quality. This data is first aligned to the reference genome and stored in a sequence alignment map (SAM) or binary alignment map (BAM) file using a sequence alignment algorithm (Li et al. 2009) (Fig. 2.1). A number of algorithms have been developed for this purpose, using Burrows-Wheeler Transformation (BWT) techniques (*e.g.*, Bowtie/Bowtie 2 (Langmead and Salzberg 2012), BWA/BWA-SW (Li and Durbin 2010)) and/or Smith-Waterman (SW) dynamic programing (*e.g.*, SHRiMP/SHRiMP2 (David et al. 2011; Rumble et al. 2009)).

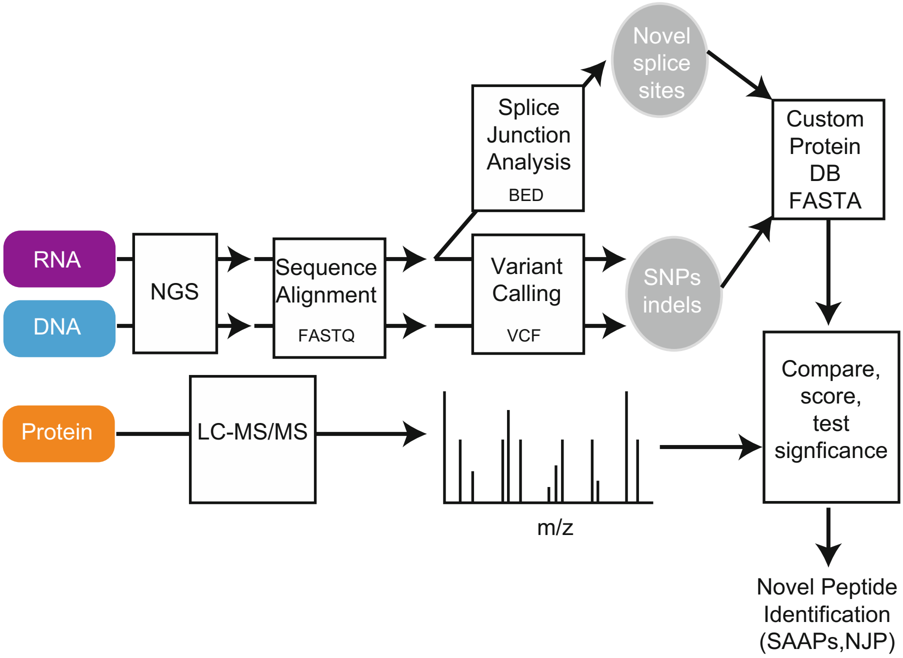


Fig. 2.1 Proteogenomic overview

2.2 Variant Identification Using Proteogenomics

2.2.1 Single Nucleotide Polymorphisms (SNPs)

Following alignment of the DNA or RNA sequence, subsequent variant calling, filtering and annotation can be completed. Variants are found through the identification of small deviations between the experimental and the reference genome (Figs. 2.1 and 2.2). These variants may be disease drivers, or mutations having little to no functional impact. Several programs have been developed specifically for the purpose of variant calling and each produces a list of variant positions stored in a Variant Call Format (VCF) file (Danecek et al. 2011).

A primary challenge with SNP variant calling is in identifying "true" variants and filtering out those due to errors in sequencing or alignment (Nielsen et al. 2011). Informatics packages have been developed for variant calling, including the popular Genome Analysis Toolkit (GATK) (McKenna et al. 2010) and VarScan (Koboldt et al. 2012). Indel mutation identification presents an additional set of complications, because it requires a more sophisticated approach to gapped alignment and paired-end sequence inference. Pattern growth approach software (*e.g.*, Pindel (Ye et al. 2009)), baysian-based algorithms (*e.g.*, Dindel (Albers et al. 2011)) and the variant calling algorithm GATK (McKenna et al. 2010) have all been refined for accurate indel identification (Neuman et al. 2013).

Following variant calling, filtering and annotation are common steps for isolating variants most likely to contribute to the pathology of interest. Although quality cutoffs for variant identification should always be employed, additional filtering becomes less important in proteogenomic analysis because proteomic data can be leveraged for variant validation.

2.2.2 Single Amino Acid Polymorphisms (SAAP)

Identifying variants that are expressed at the protein level presents a non-trivial informatics challenge in that mass spectrometric identification of peptide sequences is dependent upon the inclusion of that sequence in the protein database. Protein sequence database searching algorithms such as X!Tandem (Craig et al. 2005), Mascot (Perkins et al. 1999) and MSGF+ (Granholm et al. 2014) match the MS/MS spectra against a list of candidate peptide sequences and score the similarity of

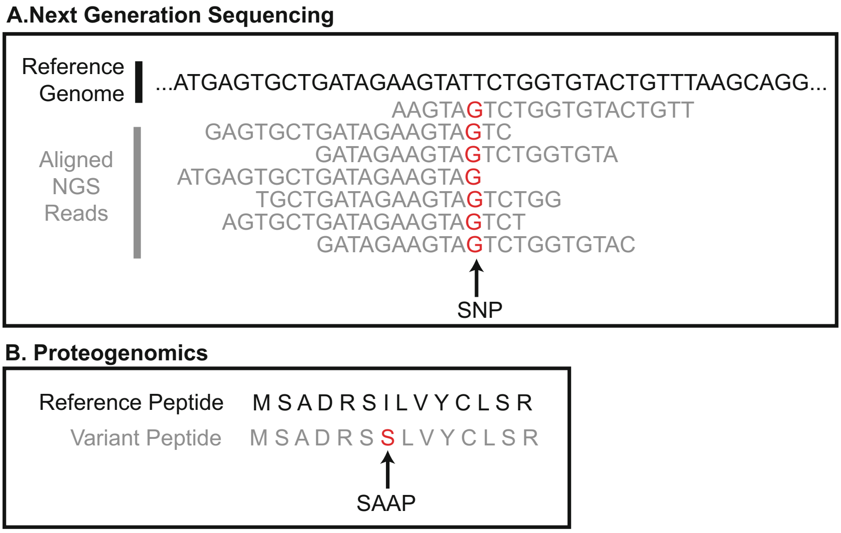


Fig. 2.2 Single nucleotide polymorphism and single amino acid polymorphism identification

a theoretical or library spectrum to the acquired spectrum based on mass. Databases with missing sequences will fail to identify these peptides in the MS/MS data and ideally, the protein database would contain all proteins present in the sample with minimal irrelevant sequences (Fig. 2.1).

Therefore, in order to identify single amino acid polymorphisms (SAAPs) occurring from non-synonymous genomic SNPs, one must create a protein sequence database that incorporates the sequencing data to contain corresponding SAAPs. These changes are integrated into the protein sequence data by first modifying the genomic reference sequence to include SNPs in the genome and/or transcriptome (Fig. 2.2a) and then completing an *in silico* protein translation of the modified sequences to attain a list of peptides containing SAAPs (Fig. 2.2b).

2.2.3 Bioinformatics Tools for Creating SAAP Protein Sequence Databases

Several tools have been developed to create NSG-integrated databases containing potential variant SAAPs. With the inclusions of these novel peptide sequences in the database, variant peptides can be identified from MS/MS data.

These tools include:

- **QUILTS**: Open source tool that incorporates SNPs from either DNA sequencing or RNA-Seq and allows for up to two variant VCF input files to accommodate cancer studies which require both germline and somatic (cancer specific) SNP options. QUILTS then creates a FASTA-formatted protein sequence database that can be used by common database searching algorithms (Ruggles et al. 2015). *quilts.fenyolab.org*
- **customproDB:** R package developed for customized protein database construction using SNPs and indels from RNA-Seq data. The output is also a FASTA-formatted sequence file (Wang and Zhang 2013). *www.bioconductor.org/packages/release/bioc/html/customProDB.html*

2.3 Alternative Splicing and Gene Annotation

Coding of novel gene regions and alternative splicing provides additional biological complexity. The advent of RNA-Seq has shown alternative splicing to occur in over 90 % of human genes (Pal et al. 2012), emphasizing the role of diverse protein isoforms in cellular function. RNA-Seq analysis provides information on splice junctions (intron / exon boundaries) present in a given sample, providing insight into both normal gene annotation and novel expression. Splice sites are identified following sequence alignment using splice-alignment software such as TopHat (Kim et al. 2013), BLAT (Fonseca et al. 2012) and MapSplice (Wang et al. 2010) (Fig. 2.1).

Comparing intron / exon boundaries identified through NGS to known junction boundaries can identify novel splice sites, including unannotated alternative splicing (two known exons) (Fig. 2.3a), partially novel splicing (one known exon) (Fig. 2.3b) and completely novel splicing (no known exons) (Fig. 2.3c) (Ruggles et al. 2015; Mertins et al. 2016). Hundreds of thousands of novel splice sites can be identified by one RNA-Seq experiment, but the fraction of functional versus “spurious” splicing requires additional information to be determined. Since *ab initio* methods for the identification of novel splice sites are limited (Barash and Garcia 2014), the validation of splice-junctions requires peptide evidence spanning these intron / exon boundaries.

2.3.1 Novel Splice Junction (NSJ) Peptides

As with SAAPs, NSJ peptide identification relies on the construction of a comprehensive protein database incorporating alternatively spliced isoforms and novel expression as coded in the transcriptome. These databases should contain all possible NSJ peptides in the sample to insure corresponding peptide identification from tandem MS analysis. Approximately one quarter of

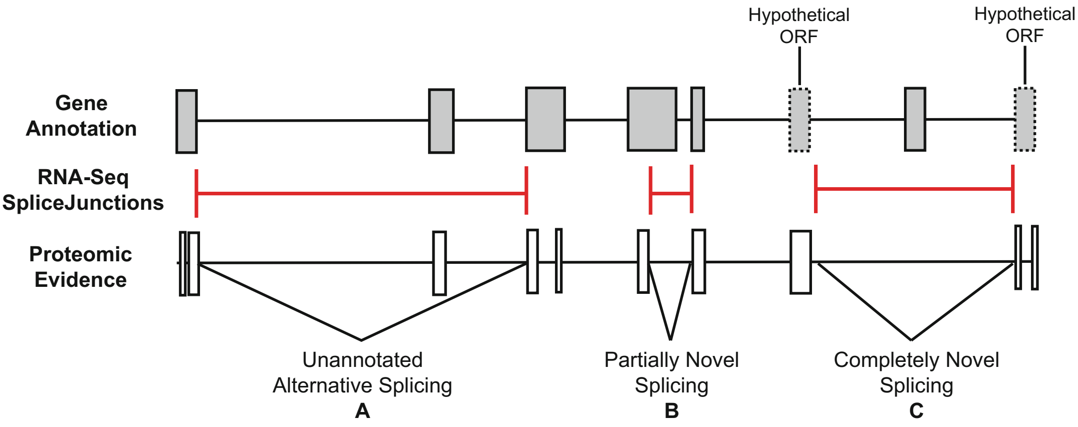


Fig. 2.3 Proteogenomic gene annotation

peptides cross a splice junction in humans, and these are particularly useful for intron / exon boundary and splicing verification. The identification of novel splice sites is most frequently used in:

- Improving gene annotation
- Cancer studies, where alternative splicing and novel expression have been reported to effect disease progression (Ning and Nesvizhskii 2010)

Gene annotation is the process of identifying genes and determining gene function. Prior to NSG, the identification of protein coding regions was done using comparative sequence analysis and gene prediction algorithms, both of which have inherent limitations. These limitations include difficulties in identifying gene start and stop sites and translational reading frames (Brent 2008), difficulty identifying splice boundaries (Reese et al. 2000), and issues in determining boundaries of short and overlapping genes (Warren et al. 2010). RNA-Seq has addressed most of these limitations, but studies have shown that many transcripts show no evidence of protein translation (Clamp et al. 2007; Eddy 2001). Proteogenomics is able to fill this gap by using MS-based proteomics in combination with RNA sequencing to verify gene coding regions and novel splice junctions.

2.3.2 Bioinformatics Tools for Identifying NSJ Peptides

A frequently applied method for proteogenomic gene annotation is the use of a six-frame translation of the DNA sequence of interest as the protein sequence database for the MS/MS peptide search, removing all bias based on the established genome annotation. This method is able to validate existing gene models and start / stop sites, and can also identify novel open reading frames (ORFs) (Fermin et al. 2006; Gupta et al. 2007; Kalume et al. 2005). Two limitations to this method are:

- The inclusion of a six-frame translation considerably increases the search space thereby reducing search sensitivity
- Splicing information cannot be determined, only intron / exon boundaries.

Tools for six-frame translation database searches include:

- **Peppy:** A Java-based software that searches a given six-frame translation database, returning peptide identifications at a user-specified false discovery rate (Risk et al. 2013). *http://geneffects.com/pepp*

- **PIUS (Peptide Identification by Unbiased Search):** Online tool that identifies peptides through a spectral match search of high-throughput MS/MS data using a six-frame translation database (Costa et al. 2013).

An alternative to a six-frame translation search is to use RNA-Seq derived splice junction data to identify novel alternative splicing in addition to unannotated ORFs. This requires more sophisticated informatics tools, which incorporate cases of unannotated alternative splicing (Fig. 2.3a), splicing at a novel intron / exon boundary (Fig. 2.3b), and splicing of novel, hypothetical open reading frames (Fig. 2.3c) to the genomic reference sequence. A Browser Extensible Data (BED) file, containing information on the location of these junctions, is created by most RNA-Seq alignment algorithms and used in the sequence modification step. The protein database is then created using an *in silico* protein translation of these modified sequences to obtain a full NSJ peptide list. Translation of these splice junctions can then be verified by the identification of peptide sequences bridging the transcribed intron / exon boundaries.

Tools that create NSJ protein sequence databases include:

- **QUILTS:** In addition to incorporating SNPs from NSG data to the protein sequence database, QUILTS accepts a Browser Extensible Data (BED) file containing RNA-Seq predicted splice junctions as input and creates FASTA files containing NSJ peptides corresponding to the transcriptome data (Ruggles et al. 2015). *quilts.fenyolab.org*
- **customproDB:** In addition to SNP-based protein database creation, customproDB creates FASTA database files using a putative junction BED file (Wang and Zhang 2013) *www.bioconductor.org/packages/release/bioc/html/customProDB.html*

2.4 Coordinated Gene and Protein Expression

In addition to facilitating the identification of SAAPs and NSJ peptides, proteogenomics can also support coordinated expression analysis based on genomic location. Copy number variation (CNV), defined as large (>1 kb) genomic deletions / duplications, can be derived from whole genome and exome sequencing. CNVs often result in gene dosage effects in multiple genes and have been shown to play a significant role in genetic variation and disease (Iafrate et al. 2004). Most methods for CNV detection can be categorized into two types: pair end mapping (PEM) methods and depth of coverage (DOC) methods. The more popular DOC algorithms such as SegSeq (Chiang et al. 2009) and CNV-seq (Xie and Tammi 2009) align reads on the genome and calculate read counts using sliding bins, which are further processed to determine a normalized copy number (Duan et al. 2013).

At the transcript level, differential gene expression is determined using methods that use read coverage to quantify transcript abundance. For example, RPKM (reads per kilo base per million mapped reads) and FPKM (fragments per kilobase per million) are commonly used methods to quantify normalized expression of a gene (Marioni et al. 2008) and many programs have been developed for the subsequent determination of differential gene expression, these include Cuffdiff (Trapnell et al. 2013), edgeR (Robinson et al. 2010), and DESeq (Anders and Huber 2010).

Proteogenomic tools have been developed that allow for coordinated expression analysis across data types, by converting proteomic location to genomic coordinates. This mapping allows researchers to analyze expression based on genomic location, for example in large areas of gene duplication / deletion, or at the exon level, rather than requiring gene-based analysis. This is particularly useful when displaying expression levels using genome browsers (*e.g.*, UCSF genome browser, Integrative Genomics Viewer (IGV)).

Bioinformatics tools for peptide mapping include:

- **PGx:** Open-source tool that maps peptides onto their putative genomic coordinates using a user-defined reference database. The software maps many peptides simultaneously, returning a BED (qualitative) and bedGraph (quantitative) file, which can be used to then be loaded into a genome browser for visualization (Askenazi et al. 2015). pgx.fenyolab.org
- **The Proteogenomic Mapping Tool:** Java-based software that searches peptides against a six-frame translated sequence database. Output includes a file containing the genomic location of each peptide match that can be visualized using a genome browser (Sanders et al. 2011). www.agbase.msstate.edu/tools/pgm/

2.5 Conclusions

Informatics-based proteogenomic methods help to determine which genomic variants and alternatively spliced gene forms are translated, revealing their biological potential. For example, mutations and novel splice junctions that are found at the peptide level have a higher likelihood of being a driver of disease. Additionally, integrating NGS and proteomic data using proteogenomic mapping tools allows for the simultaneous analysis of gene expression, which can help to better understand the complexities of gene regulation. We expect that as NGS and high throughput proteomic techniques continue to improve, the quantity and quality of associated data will continue to rise and will demand continuously evolving bioinformatics tools for proteogenomic integration and analysis.

Acknowledgement This work was supported by funding provided by the National Institutes of Health through contract S13-068 from Leidos.

References

Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2011). Dindel: Accurate indel calls from short-read data. *Genome Research, 21*, 961–973. doi:10.1101/gr.112326.110.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*, R106. doi:10.1186/gb-2010-11-10-r106.

Askenazi, M., Ruggles, K. V., & Fenyö, D. (2015). PGx: Putting peptides to BED. *Journal of Proteome Research*. doi:10.1021/acs.jproteome.5b00870.

Barash, Y., & Garcia, J. V. (2014). Predicting alternative splicing. *Methods Molecular Biology, 1126*, 411–423. doi:10.1007/978-1-62703-980-2_28.

Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews Genetics, 9*, 62–73. doi:10.1038/nrg2220.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., & Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods, 6*, 99–103. doi:10.1038/nmeth.1276.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., & Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 19428–19433. doi:10.1073/pnas.0709013104.

Costa, E. P., Menschaert, G., Luyten, W., De Grave, K., & Ramon, J. (2013). PIUS: Peptide identification by unbiased search. *Bioinformatics, 29*, 1913–1914. doi:10.1093/bioinformatics/btt298.

Craig, R., Cortens, J. P., & Beavis, R. C. (2005). The use of proteotypic peptide libraries for protein identification. *Rapid Communications in Mass Spectrometry, 19*, 1844–1850. doi:10.1002/rcm.1992.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*, 2156–2158. doi:10.1093/bioinformatics/btr330.

David, M., Dzamba, M., Lister, D., Ilie, L., & Brudno, M. (2011). SHRiMP2: Sensitive yet practical SHort read mapping. *Bioinformatics, 27*, 1011–1012. doi:10.1093/bioinformatics/btr046.

Duan, J., Zhang, J.-G., Deng, H.-W., & Wang, Y.-P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS One, 8*, e59128. doi:10.1371/journal.pone.0059128.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics, 2*, 919–929. doi:10.1038/35103511.

Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G. S., & States, D. J. (2006). Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biology, 7*, R35. doi:10.1186/gb-2006-7-4-r35.

Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics, 28*, 3169–3177. doi:10.1093/bioinformatics/bts605.

Granholm, V., Kim, S., Navarro, J. C. F., Sjölund, E., Smith, R. D., & Käll, L. (2014). Fast and accurate database searches with MS-GF+Percolator. *Journal of Proteome Research, 13*, 890–897. doi:10.1021/pr400937n.

Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R. D., & Pevzner, P. A. (2007). Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Research, 17*, 1362–1377. doi:10.1101/gr.6427907.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics, 36*, 949–951. doi:10.1038/ng1416.

Kalume, D. E., Peri, S., Reddy, R., Zhong, J., Okulate, M., Kumar, N., & Pandey, A. (2005). Genome annotation of Anopheles gambiae using mass spectrometry-derived data. *BMC Genomics, 6*, 128. doi:10.1186/1471-2164-6-128.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*, R36. doi:10.1186/gb-2013-14-4-r36.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research, 22*, 568–576. doi:10.1101/gr.129684.111.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*, 357–359. doi:10.1038/nmeth.1923.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*, 589–595. doi:10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*, 2078–2079. doi:10.1093/bioinformatics/btp352.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research, 18*, 1509–1517. doi:10.1101/gr.079558.108.

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Mundt, F., Tu, Z., Lei, J. T., Gatza, M., Perou, C. M., Yellapantula, V., Lin, C., Ding, L., McLellan, M., Ping, Y., Davies, S. R., Townsend, R., Zhang, B., Rodriguez, H., Paulovich, A., Fenyo, D., Ellis, M., Carr, S. A., & The NCI CPTAC. (2016). Proteogenomic analysis of human breast cancer connects genetic alterations to phosphorylation networks. *Nature, 534*(7605), 55–62.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A mapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*, 1297–1303. doi:10.1101/gr.107524.110.

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the basics of NGS: From mechanism to variant calling. *Current Genetic Medicine Reports, 3*, 158–165. doi:10.1007/s40142-015-0076-8.

Neuman, J. A., Isakov, O., & Shomron, N. (2013). Analysis of insertion-deletion from deep-sequencing data: Software evaluation for optimal detection. *Briefings in Bioinformatics, 14*, 46–55. doi:10.1093/bib/bbs013.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics, 12*, 443–451. doi:10.1038/nrg2986.

Ning, K., & Nesvizhskii, A. I. (2010). The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: A preliminary assessment. *BMC Bioinformatics, 11*(Suppl 11), S14. doi:10.1186/1471-2105-11-S11-S14

Pal, S., Gupta, R., & Davuluri, R. V. (2012). Alternative transcription and alternative splicing in cancer. *Pharmacolology & Therapeutics, 136*(3), 283–294. doi:10.1016/j.pharmthera.2012.08.005.

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis, 20*, 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F., & Lewis, S. E. (2000). Genome annotation assessment in Drosophila melanogaster. *Genome Research, 10*, 483–501.

Risk, B. A., Spitzer, W. J., & Giddings, M. C. (2013). Peppy: Proteogenomic search software. *Journal of Proteome Research, 12*, 3019–3025. doi:10.1021/pr400208w.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data.

Bioinformatics, 26, 139–140. doi:10.1093/bioinformatics/btp616.

Ruggles, K. V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M. D., Clauser, K. R., Tabb, D. L., Mertins, P., Slebos, R., Erdmann-Gilmore, P., Li, S., Gunawardena, H. P., Xie, L., Liu, T., Zhou, J.-Y., Sun, S., Hoadley, K. A., Perou, C. M., Chen, X., Davies, S. R., Maher, C. A., Kinsinger, C. R., Rodland, K. D., Zhang, H., Zhang, Z., Ding, L., Townsend, R. R., Rodriguez, H., Chan, D., Smith, R. D., Liebler, D. C., Carr, S. A., Payne, S., Ellis, M. J., & Fenyo, D. (2015). An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Molecular Cellular Proteomics*. doi:10.1074/mcp.M115.056226.

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., & Brudno, M. (2009). SHRiMP: Accurate mapping of short color-space reads. *PLoS Computational Biology, 5*, e1000386. doi:10.1371/journal.pcbi.1000386.

Sanders, W. S., Wang, N., Bridges, S. M., Malone, B. M., Dandass, Y. S., McCarthy, F. M., Nanduri, B., Lawrence, M. L., & Burgess, S. C. (2011). The proteogenomic mapping tool. *BMC Bioinformatics, 12*, 115. doi:10.1186/1471-2105-12-115.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nature Biotechnology, 31*, 46–53. doi:10.1038/nbt.2450.

Wang, X., & Zhang, B. (2013). customProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics, 29*, 3235–3237. doi:10.1093/bioinformatics/btt543.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., & Liu, J. (2010). MapSplice: Accurate mapping of RNA-Seq reads for splice junction discovery. *Nucleic Acids Research, 38*, e178. doi:10.1093/nar/gkq622.

Warren, A. S., Archuleta, J., Feng, W.-C., & Setubal, J. C. (2010). Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics, 11*, 131. doi:10.1186/1471-2105-11-131.

Xie, C., & Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics, 10*, 80. doi:10.1186/1471-2105-10-80.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics, 25*, 2865–2871. doi:10.1093/bioinformatics/btp394.

3 Proteogenomics: Key Driver for Clinical Discovery and Personalized Medicine

Ruggero Barbieri, Victor Guryev, Corry-Anke Brandsma, Frank Suits, Rainer Bischoff, and Peter Horvatovich

Abstract

Proteogenomics is a multi-omics research field that has the aim to efficiently integrate genomics, transcriptomics and proteomics. With this approach it is possible to identify new patient-specific proteoforms that may have implications in disease development, specifically in cancer. Understanding the impact of a large number of mutations detected at the genomics level is needed to assess the effects at the proteome level. Proteogenomics data integration would help in identifying molecular changes that are persistent across multiple molecular layers and enable better interpretation of molecular mechanisms of disease, such as the causal relationship between single nucleotide polymorphisms (SNPs) and the expression of transcripts and translation of proteins compared to mainstream proteomics approaches. Identifying patient-specific protein forms and getting a better picture of molecular mechanisms of disease opens the avenue for precision and personalized medicine. Proteogenomics is, however, a challenging interdisciplinary science that requires the understanding of sample preparation, data acquisition and processing for genomics, transcriptomics and proteomics. This chapter aims to guide the reader through the technology and bioinformatics aspects of these multi-omics approaches, illustrated with proteogenomics applications having clinical or biological relevance.

R. Barbieri
Department of Gastroenterology and Hepatology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

V. Guryev
European Research Institute for the Biology of Ageing, University Medical Center Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands

C.-A. Brandsma
Department of Pathology & Medical Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

F. Suits
IBM T.J. Watson Research Centre, 1101 Kitchawan Road, Yorktown Heights, New York 10598, NY, USA

R. Bischoff • P. Horvatovich (✉)
Department of Analytical Biochemistry, Research Institute of Pharmacy, University of Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands
e-mail: p.l.horvatovich@rug.nl

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_3

Keywords

Genomics • Proteomics • Mass spectrometry • Proteogenomics • DNA/RNA next generation sequencing

3.1 Introduction

Genome sequencing technology aims to reveal the nucleotide sequence of the genome and stage-specific transcriptome states across different cells and tissues. The proteome is defined as "the protein complement of the genome". Proteins are the product of the translated part of the genome and transcriptome. Proteins are biologically active molecules, while genomes and transcriptomes, besides exerting a regulatory role, hold information on possible protein primary sequences that the cells of an organism can express and use to fulfill their molecular activities and biological functions.

Sequencing DNA or mRNA requires an analytical system that distinguishes precisely between the nucleobases cytosine, guanine adenine (DNA, RNA), thymine (DNA) and uracil (RNA). Combinations of these five bases represent a much simpler chemical system compared to the chemical space spanned by the twenty amino acids and their possible chemical modifications, e.g. through post-translational modifications that form the proteins (Chuh and Pratt 2015; Walsh et al. 2005; Markiv et al. 2012; Bischoff and Schlüter 2012). This larger and more diverse chemical space and the currently available peptides and protein sequencing technologies are not sensitive and powerful with respect to sequencing length compared with current state-of-the-art DNA and RNA sequencing technologies. Additionally, the information content at the genomics and transcriptomics level can be easily amplified, but no such technology exists for proteomics. The main difference between mainstream next generation sequencing technology and shotgun bottom-up LC-MS/MS proteomics is that the former provides hypothesis-free *de novo* sequencing data, from which the sequence of base pairs can be determined without prior information. However, proteomics analysis determines the primary amino acid sequence from an often incomplete list of fragment ions resulting from the fragmentation of peptides constituting the initial protein. Not all the obtained fragment ion or MS/MS spectra are suitable for a hypothesis-free *de novo* sequence determination of the fragmented peptide. Therefore the most popular approach to analyze shotgun LC-MS/MS spectra are based on targeted database search (DBS) algorithms, which uses a list of protein sequences that are expected to be present in the analyzed sample. This approach is therefore hypothesis-driven and the success of the identification relies on the accurate prediction of the protein sequence that is expected to be present in the sample. In order to provide accurate sequence information, the proteomics community uses sequences assembled by consortia or large groups that have been quality-controlled either manually (SwissProt) or computationally (TrEMBL and Ensembl). The definition of canonical sequences according to the most widely used UniProtKB/SwissProt database (Consortium 2015) is:

1. The protein sequence of all the protein products encoded by one gene in a given species is represented in a single entry to reduce protein sequence redundancy
2. The canonical sequence includes the protein sequence that has the highest occurrence
3. The canonical protein sequence shows the highest similarity to orthologous sequences found in other species
4. The length of the sequence or amino acid composition allows the clearest description of protein domains, isoforms, polymorphisms and post-translational modifications (PTMs)
5. In the absence of any other information the longest sequence is chosen

For organisms – amongst them humans, for which the genome sequence is completed – the protein sequence derived from genome translation is used, unless there is clear evidence that a different polymorphism is more frequent at a given position.

It is clear from the definition of the canonical sequence that it represents an average sequence of the proteome, but it cannot be used to detect peptides specific for low frequency variants or new variants. The proteogenomics approach performs next generation sequencing of a genome and/or transcriptome in the same sample and composes the protein sequence used during DBS of peptide and protein inference. This composition is not a trivial task and gene models that predict the translation of genomics sequences into proteins are used. In the early days of the genomics era, proteogenomics was defined as a description of "studies in which proteomic data are used for improved genome annotation and characterization of the protein-coding potential" (Nesvizhskii 2014; Menschaert and Fenyo 2015; Bischoff et al. 2015). Therefore in the early days the proteomics dataset helped to provide accurate genome annotation. Nowadays it is more frequent to use the genomic sequence information to obtain sample, or in clinical research patient-specific protein sequence information and predict which protein forms are present in a given sample. Therefore, proteogenomics data analysis allows better and more accurate protein identification and better reflects the biological processes that are active in the cell and/or tissue of the analyzed sample. Since a high quality patient specific database is used for peptide and protein identification, proteogenomics enables a personalized approach to identify patient specific molecular heterogeneity and novel patient phenotypes within a disease. Furthermore, it allows discovery of biomarkers for its specific diagnosis, as well as the discovery of new drug targets that allow more precision and personalized treatment. Importantly, proteogenomics analysis has become more affordable by the reduction of sequencing costs, which has enabled the generation of more precise information of clinical samples, and thus patient specific proteomes, when compared to mainstream proteomics analysis using public databases.

This chapter has the primary aim to provide an overview of the main characteristics of data obtained with next generation sequencing technology combined with the shotgun LC-MS/MS proteomics approach, to describe the key data processing steps and the integrated data interpretation of these two molecular layers. The chapter is intended for readers interested in the data analysis and interpretation of one or both -omics fields with the ultimate goal to perform a proteogenomics analysis. Best practice in data acquisition, data processing approaches and challenges with respect to data and analysis tools will be thoroughly discussed.

3.2 RNA and DNA Sequencing

3.2.1 Genomic Sequencing Technologies

The translated protein sequence can be deduced from full genome, exome and transcriptome sequencing data but the most widely used approach is polyadenylated transcriptome sequencing (RNA-Seq). Figure 3.1 summarizes the starting molecular level (DNA, mRNA), and the applied protocols and factors that should be taken into account during sequencing. Sequencing the full genome costs an order of magnitude more than sequencing exomes or transcriptomes. For DNA sequencing, the two main options are Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES). While the first gives a complete overview of variations in the genome, the second covers the coding part of the genome (exome), which accounts only for several percent of the complete genome. For these reasons WES could be a good choice as source of genetic information for a proteogenomics approach. Sequencing polyadenylated mRNA has the advantage that the majority of the transcripts were already processed by the splicing machinery, resulting in a high fraction (>90 %) of mature transcripts with spliced introns, which provide the highest quality of sequence information to predict the sequence of the translated proteome. The other alternative is the removal of highly abundant ribosome mRNA with special kits (Ribo-zero kits), which enables the user to sequence the complete transcriptome that contains the translated and non-translated transcripts.

There are different technologies for transcriptome and genome sequencing. In the early days of sequencing technology, DNA sequencing was used to obtain the first complete genomes. The most important milestones were deciphering of the genome of the bacteriophage ΦX 174 (Sanger et al. 1977) (first complete genome) by Frederick Sanger and the human genome by the Human Genome Project (Lander et al. 2001). The technique developed by Frederick Sanger was the first to be automated and is considered as "first generation" of sequencing technology. The "second generation" (also called Next Generation Sequencing) started with MPSS (Massively Parallel Signature Sequencing) from Lynx Therapeutics and was characterized by cheaper, faster and more efficient sequencing, which led to the acquisition of an enormous amount of genomic information. Nowadays, the most widely used sequencing technology is short-read based sequencing, with the Illumina HiSeq sequencing machines. Figure 3.1 shows the standard protocol for DNA sequencing, which technology has not only drastically reduced the time necessary for sequencing but also the cost of each analysis run, leading to complete transcriptome sequencing in a matter of hours. Typical fragment sizes range from 100 bps up to 600 bps. Fragments are then read from one (single-end) or both sides (paired-end) up to 250 bps.

There are various options when preparing samples for the sequencing run(s), based on different protocols that focus on different types of transcripts or different ways to analyze them. While at first only the coding messages of the transcriptome were sequenced, through a selection of polyadenylated transcripts (the mRNAs that are most likely to be translated into proteins), the growing interest in the non-coding transcriptome has led to a different approach, where only the major non-coding RNA type, the ribosomal RNA (rRNA), is depleted. This protocol is defined as rRNA depletion or the Ribo-zero approach and is achieved with special ribosomal mRNA removal kits. For a proteogenomics approach, it is often considered a good choice to use the polyadenylated mRNA protocol and thus focus only on protein-generating transcripts, to minimize the error rate and provide the most accurate protein sequence information that is supposed to be expressed in the cells of the target organism. Conversely, the rRNA-depletion protocol retains long non-coding RNAs (or lncRNAs) and other non-polyadenylated transcripts, which are thought to have a regulatory function. However, there is growing evidence that some of the lncRNAs might be translated. lncRNAs are

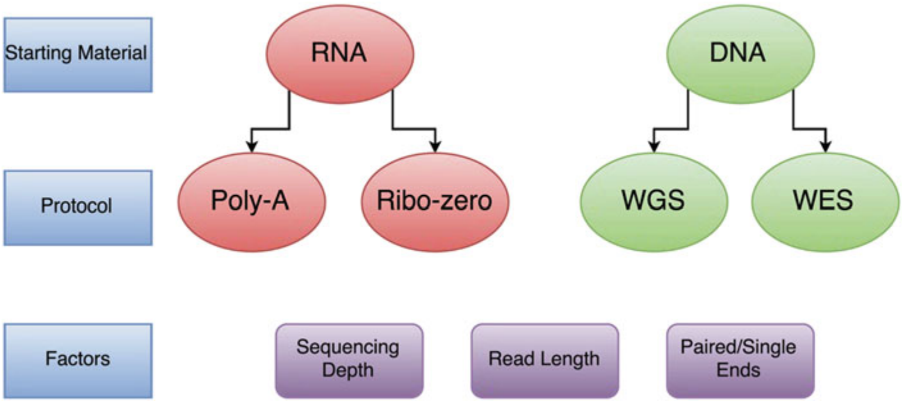


Fig. 3.1 Chart showing the different molecular classes that can be sequenced using next generation sequencing (RNA *red*, DNA *green*), showing the starting material, the sequencing protocols and experimental factors (*purple rectangles*) that can be set by the user. *WGS* represents whole genome sequencing and *WES* represents whole exome sequencing

lineage-specific and it is hypothesized that they show similar characteristics as evolutionary young protein coding genes (Ruiz-Orera et al. 2014). Proteomics identification of lncRNA from large public database such as the PRIDE repository (http://www.ebi.ac.uk/pride/archive/) showed high FDR rates of translated lncRNA sequences and therefore the results should be taken with care (Volders et al. 2013, 2015).

In addition to the decision of what should be sequenced, there are various factors that should be taken into account when designing an RNA or DNA-sequencing experiment. For example, the sequencing depth or the number of reads for each sample determine the quality of data and influence important properties such as the quality of the alignment to a reference genome, the number of identified sequence variants that differ from the reference genome and affect the reliability of quantifications. The optimal sequencing level should be determined based on the aim of the experiment. However, it is obvious that a complex sample (for example from biopsies that typically contain different types of cell from different tissues) requires higher sequencing depth when compared to a simpler sample consisting of one cell or tissue type.

In a similar fashion, the length of the reads may have a consistent effect on the quality of the post-sequencing alignment to a reference genome and thus the ability to correctly determine the transcripts structure and amount. Longer reads tend to minimize the effect of sequencing errors and capture splicing events or multi-nucleotide deletions and insertions more efficiently. On the other hand, if the intention is only to quantify the amount of transcript(s) present, short reads (such as 50 bp) may be sufficient, leading to reduced cost and analysis time.

Sequencing can be performed with two approaches concerning the reading directions of 500 base pair transcript fragments, these are known as single and paired-ends. The effect of longer reads is magnified when paired-end reads are used. There are sample preparation kits that cannot discriminate whether a sequence is read in forward or reverse direction (*e.g.*, TruSeq from Illumina) and there are kits that can deliver this information (*e.g.*, BioO Scientific's NextFlex). When strandedness information is lacking, it is still possible to predict from which strand the reads originate by exploiting the unique sequence of introns and exons of each transcript. Information on the exact sequenced strand is important when identifying variants, as each strand may carry a different allele (a different base in the corresponding position on each strand), could be coding (contain the translated amino acid sequence) or template (contains the complementary nucleotide base sequence) and strands originate from maternal and paternal chromosomes. In paired-end sequencing the sequence is first read in one direction and then from the opposite direction having around 500 bps of distance between the two ends (Fig. 3.2). It is important to note that one read covers a relatively small part of a fragment, but taking the fragment length into account for alignment to a reference genome provides more accurate alignment than with single-end reading. The two reads in paired-end sequencing are also called a "mate-pair". Paired-end reads provide more accurate

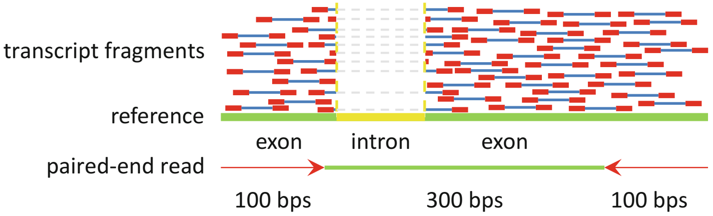


Fig. 3.2 Paired-end sequencing of fragments using the Illumina sequencer and alignment of the sequence reads to a reference genome matching the reads at both ends and taking the length of fragments into account. Reference genome sequence part with two exons and one intron is shown

data when trying to detect large modifications of the genome and transcriptome, like large insertions, deletions and translocations (also called Copy Number Variations, or CNVs). Single-end reads are less potent in this respect, but their generation is less expensive and requires shorter analysis time. Single-end sequencing is a suitable approach when transcript quantification is the sole aim of the experiment.

3.2.2 Sequencing Technology

A typical protocol for short read sequencers is shown in Fig. 3.3 and is composed of the following steps:

1. DNA/RNA extraction and isolation from sample to retrieve the DNA/polyadenylated mRNA fraction. Extraction can be performed with different protocols using chemical approaches such as phenol-based extractions, direct lysis of DNA and RNA strands or using a mechanical approach such as centrifugation trough molecular filters of defined size and recovery of the nucleic acids with magnetic beads coated with DNA/RNA binding molecules.
2. This is followed by fragmentation of the extracted DNA and mRNA to obtain shorter pieces that can be efficiently sequenced. Fragmentation can be achieved with different methods. The most common is using enzymes that cut the nucleic acids randomly (though the sites where each enzyme cuts is known, it is therefore possible to infer the fragment distribution), by sonication (use of high amplitude sound waves to break DNA and RNA strands) or by intense heating. These steps are followed by selecting fragments of desired length, which is usually performed using a size exclusion gel electrophoresis. There are alternative approaches for size selection such as using magnetic beads by adjusting the concentration of the nucleic acid-binding agents present on the surface of the beads and thus selecting shorter or longer fragments. Extraction, fragmentation and fragment selection with a desired size are often performed by standardized protocols using commercial kits such as the widely used TruSeq Sample Preparation Kit from Illumina.

   In the case of mRNA analysis, transcript fragments are reverse-transcribed into cDNA, which turns the mRNA sequence into a DNA sequence (Fig. 3.4). Adapters of 6–8 nucleotides in length are ligated to each end of the fragments, which permits them to be immobilized on the surface of a flow-cell, which is a container where the sequence amplification and sequencing reaction take place. The adapters are complementary to primers already present and fixed on the surface of the flow-cell where they act as anchors when a transcript is fixed on the surface. Adapters may contain a short signature which is unique for each sample, and is called “barcode” (4–12 nucleotides long, with unique sequence for each sample). This allows multiple samples to be sequenced at the same time.

3. Polymerase-based amplification takes place and creates clusters of clones of the same fragment in a limited area called “spot”. Fragments are flexible and can bend in a way that the “free-end” of the adapter binds to another immobilized primer on the cell-flow surface. The polymerase can still bind to the immobilized primer and produce the second strand for each fragment. Due to this behavior this step is also called “bridge amplification”. This is necessary to create a cluster that can provide a signal strong enough to be measured by the light-sensitive sensor of the sequencer.
4. At this point, it is possible to start sequencing. To do so specially modified nucleotides called “labeled reversible terminators” are used. Four terminators are needed for each base (Adenine, Cytosine, Thymine and Guanine). Each is labeled with a different fluorescent fluorophore group, a light-sensitive molecule that will emit light at specific wavelength (red, green, blue and yellow) when excited by lasers of different wavelengths. The flow-cell is made of glass allowing the emitted fluorescence to be detected by a photo-sensitive

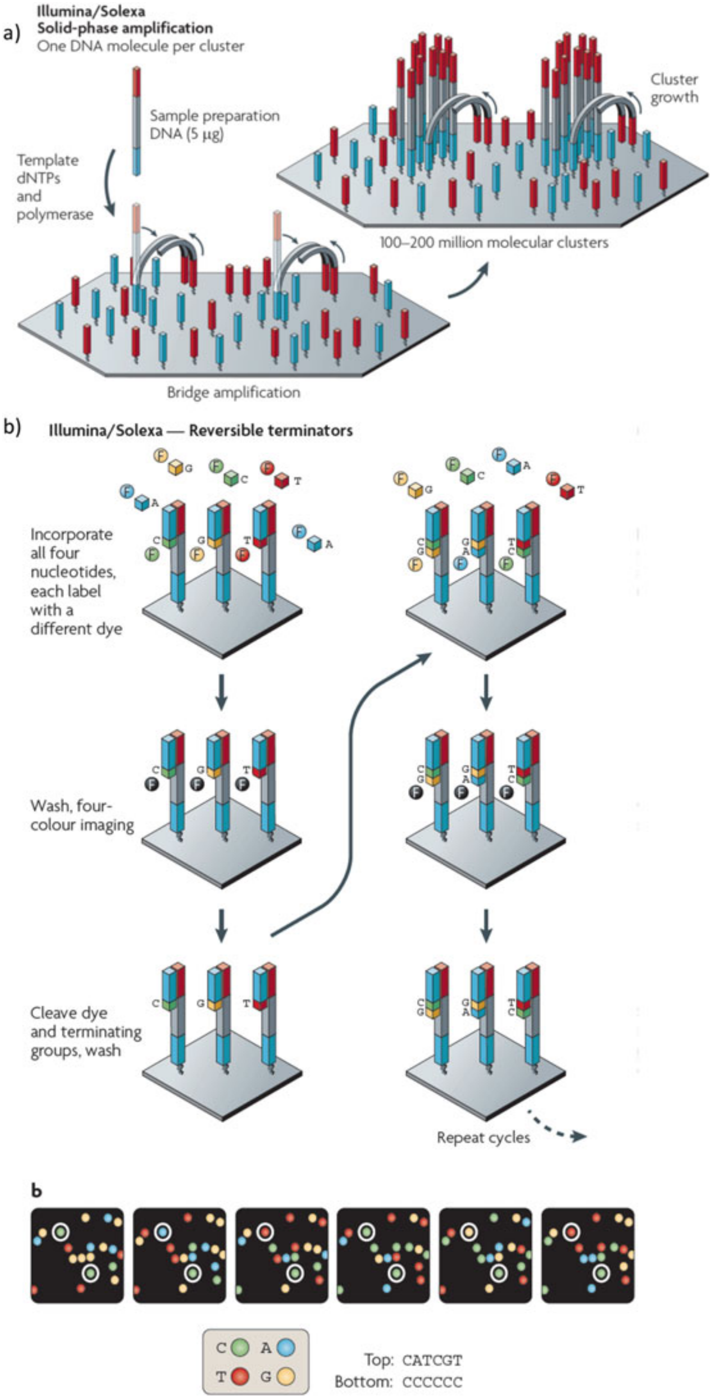


Fig. 3.3 Main step of the library preparation (**a**) and of the DNA/mRNA fragment sequencing (**b**). Further details are described in details in the main text (Figure adapted with permission from (Metzker 2010)) Copyright (2010) Nature Publishing Group

detector. Terminators stop the polymerase chain reaction as well without requiring an extra chemical reaction with this purpose that led to the name of "terminators". Reading the intensity of the emitted fluorescence at the three wavelengths emitted by the 4 terminators allow to "read" which base was added by the polymerase. Following fluorescence measurement a chemical reaction is performed to cleave the dye and the terminator group from

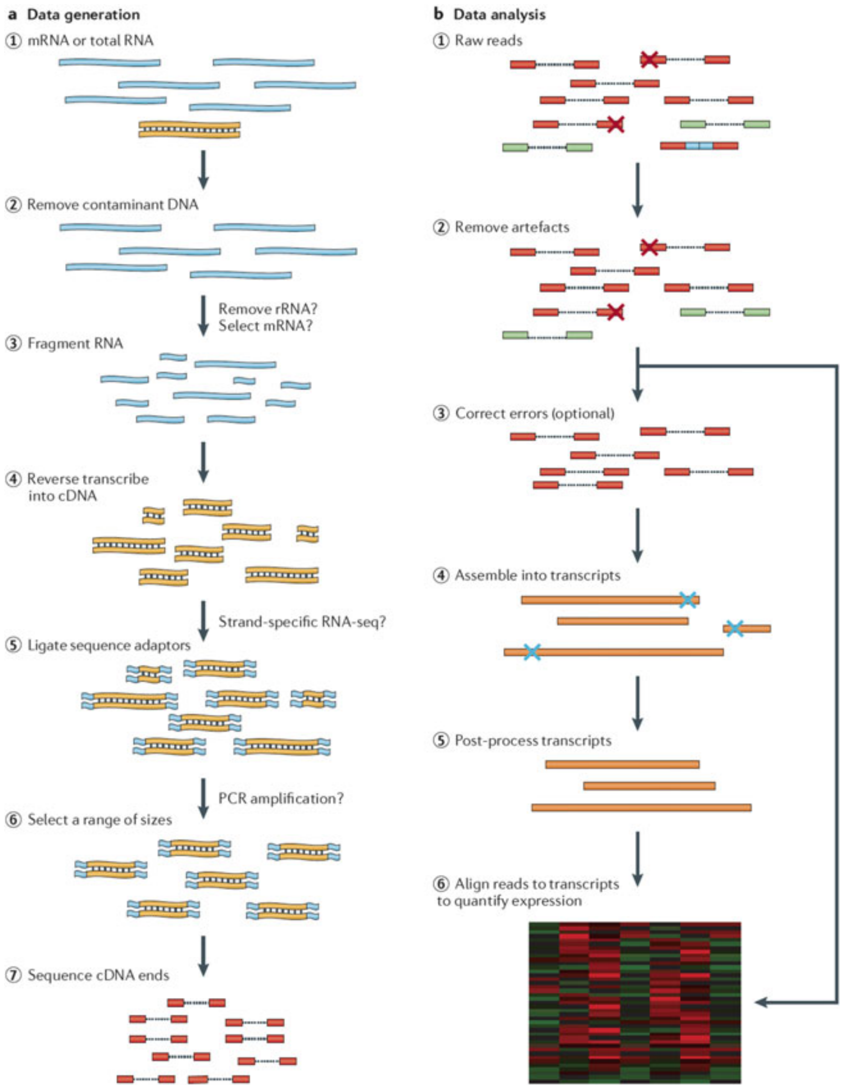
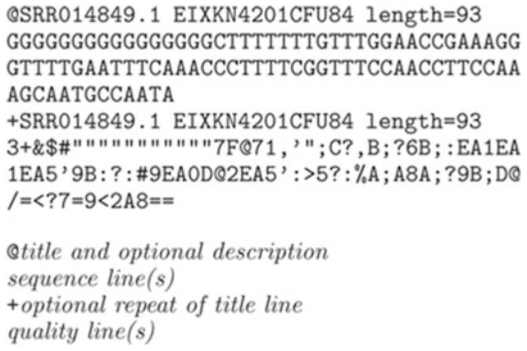

a Data generation
① mRNA or total RNA
② Remove contaminant DNA
Remove rRNA?
Select mRNA?
③ Fragment RNA
④ Reverse transcribe into cDNA
Strand-specific RNA-seq?
⑤ Ligate sequence adaptors
PCR amplification?
⑥ Select a range of sizes
⑦ Sequence cDNA ends
b Data analysis
① Raw reads
② Remove artefacts
③ Correct errors (optional)
④ Assemble into transcripts
⑤ Post-process transcripts
⑥ Align reads to transcripts to quantify expression

Fig. 3.4 Scheme of DNA and mRNA sample preparation for sequencing (*left column*) and pre-processing of raw data (*right column*) (Figure is reproduced with permission from (Martin and Wang 2011). Copyright (2011) Nature Publishing Group)

the incorporated nucleotide allowing to continue the polymerase chain reaction. This cycle is then repeated for a fixed number of times (determined by the read length), which is typically 100 or 125 in an Illumina short read sequencer.

The sequencer uses internal software to transform the measured raw fluorescence information to base pairs and includes parameters that reflect the quality of the reads. The measured base sequence is collected and saved in FastQ format (Fig. 3.5). FastQ is a simple, text based format, composed of 4 parts per entry (read): the first line starts with an "@" symbol and is an identifier of the read, which may include various kinds of information such as the length of the read, a batch ID and a read individual ID; the second part is the read itself, which may occupy more than one line depending on its length; the third part is a single line of comment starting with a '+' symbol and which may repeat the first line, report additional



```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)
```

Fig. 3.5 Example of a sequence read of transcript fragment in FastQ format (Adapted with permission from Cock et al. (2010). Copyright (2010) Oxford University Press)

information or left blank; the last part is a string of symbols, one for each letter in the read sequence. These symbols encode for numbers that represent an estimate of the quality of the resulting sequence. The numbers are calculated through a $-\log_{10}$ (estimated error rate) where the estimated error rate is the probability that the letter in a single position is correct.

The described Illumina short read sequencing technology has the advantage to deliver large amounts of sequencing data but the relatively short sequencing length of 100/125 base pairs, which information is insufficient to reconstruct large repeated regions in the genome and to reconstruct the exact transcript profile for genes with a high number of splicing events. The obtained data require significant efforts to reconstruct the transcriptome through a complex bioinformatics pipeline. Third generation sequencers are emerging, such as the PacBio sequencing technology, which provide much longer sequence reads, up to tens of thousands of consecutive bases, an approach that should resolve these issues.

3.2.3 Bioinformatics Processing of Raw Data for Proteogenomics Application

The obtained raw short sequence read data in FastQ format is processed with complex bioinformatics workflows. A workflow typically consists of the following steps:

1. Quality assessment of the reads and trimming (removing) of low quality reads
2. Assembly of short reads and alignment to the reference genome
3. Variant calling and transcript quantification
4. Prediction of translated protein sequence by finding open reading frames (ORFs) and stop codons and saving the results in protein sequence Fasta file format

The obtained protein sequence information is then used for protein and peptide identification using DBS of LC-MS/MS proteomics data, while the quantitative transcript profile is used to determine the differentially expressed transcripts in a group of samples, such as controls and samples from different stages of disease. Different tools are available for each processing step. The alignment of the raw data to the reference genome can be replaced with *de novo* hypothesis-free transcript assembly. Bioinformatics processing and the subsequent statistical analysis is an error-prone process and the quality of the obtained results should be thoroughly assessed. Each tool in the bioinformatics workflow makes different assumptions which are based on different mathematical models and algorithmic approaches, which in turn tend to capture only a part of the biological significance contained in the data. With respect to proteogenomics, the best performance assessment is to check the number of identified peptides and proteins. This assessment can be performed for different workflows built from different tools and parameters. The sample preparation protocol and the bioinformatics workflow to process RNA-Seq data is presented in Fig. 3.4:

1. The quality assessment and trimming is performed with the FastQC (Patel and Jain 2012), FastX Toolkit (Pearson et al. 1997) and Trimmomatic tools (Bolger et al. 2014) which provide a quality control report in html format for each raw FastQ files. Obviously trimming is performed only if the quality-control reports

indicate that this is necessary due to low quality of the sequence. A very common case is a drop in quality in the final part of the read due to the degradation of the efficiency of the chemical reaction of the sequencing process. These lower quality bases are generally removed at this step. The FastQC, FastX Toolkit and Trimmomatic tools are easy to use and require low computation power.

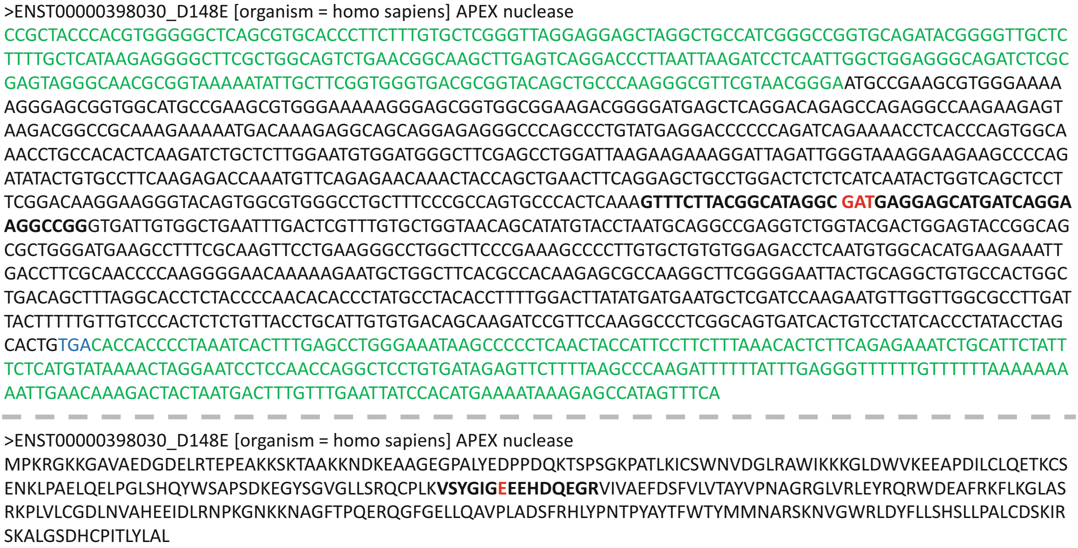
2. The trimmed FastQ files are either aligned to the reference genome using aligner tools such as STAR (Dobin et al. 2013) or Tophat2 (Kim et al. 2013). The output is then the alignment of the reads to the reference genome and the result is stored in a Binary Alignment/Map (BAM) file format. A BAM file is a compressed or binary version of a SAM or Sequence Alignment/Map format file. The SAM format follows precise specifications (see details in Lee et al. (2009)), which give the format a fixed scheme and defines where a read maps on a reference genome/transcriptome. It is composed of several lines of TAB separated fields in a fixed order, preceded by a header that gives general information on the alignment. The other option is to perform *de novo* assembly of the short reads without the use of a reference genome. This task is typically performed with tools such as ABySS (Simpson et al. 2009) and Trinity (Grabherr et al. 2011). *De novo* assembly is a computationally intensive task, as the tool needs to calculate several possible combinations of reads (grouped together in "contigs"). However reference genomes or transcriptomes are not perfect, they do contain errors, and the use of a reference genome also restricts the possibility to discover novel transcripts. Using a reference genome is a conservative choice, and can be sufficient when the analysis does not have the goal to attempt to capture all the possible transcripts in a sample or aims for maximum reliability of the assembled sequence of the transcripts.
3. The BAM file is processed by an assembler tool, which has the aim to identify the full transcript constitution in the measured sample and estimate the amount of each transcript. Commonly used transcriptome assemblers include genome reference-guided tools, such as Cufflinks (Trapnell et al. 2010), and reference-free or *de novo* transcriptome assemblers, for example Trinity (Grabherr et al. 2011). In addition a BAM file can be used as input for a genomic viewer tool, such as IGV (Integrative Genome Viewer) (Robinson et al. 2011) or Savant (Fiume et al. 2010). These genome browsers can show exactly how the reads are aligned and distributed through an easy-to-use graphical user interface, which may also include peptide abundance which data is available in a proteogenomics study.
4. At this point it is also possible to discover sequence variations in the analyzed samples. This operation is performed through the use of dedicated tools, the "variant callers", such as the HaplotypeCaller algorithm of GATK (Genome Analysis Toolkit) (McKenna et al. 2010) or the SNP Caller which is part of SAMTools (Li et al. 2009). These algorithms are able to efficiently evaluate if a SNP or insertion and deletion (indel) is present at a certain position and calculate the probability of the correctness of the findings.
5. The final steps consist of prediction of transcripts that are most likely translated into proteins and obtain the corresponding protein sequence. For this operation specialized tools, such as Transdecoder are used. This tool was conceived as an additional step to the Trinity pipeline but it can also be used as a standalone program. Transdecoder accepts files in General Transfer Format (GTF), which is a text-based TAB separated scheme used to describe genomic entities, such as transcripts or genes. GTF files are normally used for annotation of the transcripts. An alternative tool is the recently developed GeneMarkS-T (Tang et al. 2015), which is an adaptation of GeneMarkS (Besemer et al. 2001), where prokaryotic-only ORF predictor implemented in GeneMarkS was modified to translate eukaryotic transcriptomes. Transdecoder output results in a Fasta formatted protein sequence list derived directly from the transcript list

used as input. Fasta is a very simple text format for biological sequences, similar to FastQ but with only two parts, an identifier line preceded by a '>' symbol and the sequence itself in amino acid or nucleotide sequence of the transcripts/ proteins. Example of fasta file format showing the nucleotide base sequence of Apex nuclease 1 gene and corresponding amino acid sequence of the translated protein highlighting single amino acid variant (SAAV) is shown in Fig. 3.6. The sample-specific predicted amino acid sequence of translated proteins is subsequently used in DBS to identify peptides and proteins in raw LC-MS/MS data and to determine the proteome constitution of the samples. After pre-processing the transcript sequence, transcript identity and quantity is obtained. The bioinformatics workflow used to process transcripts including concrete tools with the input data is shown in Fig. 3.7.

NCBI Gene Expression Omnibus (GEO) (Barrett et al. 2013) provides repositories for raw sequencing data that can be mined and reanalyzed, for example to obtain additional information on genome or transcript expression profiles of the same or similar cell and tissue that is the aim of the study.

3.3 Proteomics Analysis

As mentioned in the introduction, the most popular shotgun bottom-up LC-MS/MS based proteomics technology is not a sequencing technology, but is based on the fragmentation of protein-derived peptides. Intact large proteins cannot be fragmented efficiently and large proteins show problems for separation by liquid chromatography (LC), a step that is required to reduce sample complexity prior to analysis with



```
>ENST00000398030_D148E [organism = homo sapiens] APEX nuclease
CCGCTACCCACGTGGGGGCTCAGCGTGCACCCTTCTTTGTGCTCGGGTTAGGAGGAGCTAGGCTGCCATCGGGCCGGTGCAGATACGGGGTTGCTC
TTTTGCTCATAAGAGGGGCTTCGCTGGCAGTCTGAACGGCAAGCTTGAGTCAGGACCCTTAATTAAGATCCTCAATTGGCTGGAGGGCAGATCTCGC
GAGTAGGGCAACGCGGTAAAAATATTGCTTCGGTGGGTGACGCGGTACAGCTGCCCAAGGGCGTTCGTAACGGGAATGCCGAAGCGTGGGAAAA
AGGGAGCGGTGGCATGCCGAAGCGTGGGAAAAAGGGAGCGGTGGCGGAAGACGGGGATGAGCTCAGGACAGAGCCAGAGGCCAAGAAGAGT
AAGACGGCCGCAAAGAAAAATGACAAAGAGGCAGCAGGAGAGGGCCCAGCCCTGTATGAGGACCCCCCAGATCAGAAAACCTCACCCAGTGGCA
AACCTGCCACACTCAAGATCTGCTCTTGGAATGTGGATGGGCTTCGAGCCTGGATTAAGAAGAAAGGATTAGATTGGGTAAAGGAAGAAGCCCCAG
ATATACTGTGCCTTCAAGAGACCAAATGTTCAGAGAACAAACTACCAGCTGAACTTCAGGAGCTGCCTGGACTCTCTCATCAATACTGGTCAGCTCCT
TCGGACAAGGAAGGGTACAGTGGCGTGGGCCTGCTTTCCCGCCAGTGCCCACTCAAAGTTTCTTACGGCATAGGC GATGAGGAGCATGATCAGGA
AGGCCGGGTGATTGTGGCTGAATTTGACTCGTTTGTGCTGGTAACAGCATATGTACCTAATGCAGGCCGAGGTCTGGTACGACTGGAGTACCGGCAG
CGCTGGGATGAAGCCTTTCGCAAGTTCCTGAAGGGCCTGGCTTCCCGAAAGCCCCTTGTGCTGTGTGGAGACCTCAATGTGGCACATGAAGAAATT
GACCTTCGCAACCCCAAGGGGAACAAAAAGAATGCTGGCTTCACGCCACAAGAGCGCCAAGGCTTCGGGGAATTACTGCAGGCTGTGCCACTGGC
TGACAGCTTTAGGCACCTCTACCCCAACACACCCTATGCCTACACCTTTTGGACTTATATGATGAATGCTCGATCCAAGAATGTTGGTTGGCGCCTTGAT
TACTTTTTGTTGTCCCACTCTCTGTTACCTGCATTGTGTGACAGCAAGATCCGTTCCAAGGCCCTCGGCAGTGATCACTGTCCTATCACCCTATACCTAG
CACTGTGACACCACCCCTAAATCACTTTGAGCCTGGGAAATAAGCCCCCTCAACTACCATTCCTTCTTTAAACACTCTTCAGAGAAATCTGCATTCTATT
TCTCATGTATAAAACTAGGAATCCTCCAACCAGGCTCCTGTGATAGAGTTCTTTTAAGCCCAAGATTTTTTATTTGAGGGTTTTTTGTTTTTTAAAAAAA
AATTGAACAAAGACTACTAATGACTTTGTTTGAATTATCCACATGAAAATAAAGAGCCATAGTTTCA
```

```
>ENST00000398030_D148E [organism = homo sapiens] APEX nuclease
MPKRGKKGAVAEDGDELRTEPEAKKSKTAAKKNDKEAAGEGPALYEDPPDQKTSPSGKPATLKICSWNVDGLRAWIKKKGLDWVKEEAPDILCLQETKCS
ENKLPAELQELPGLSHQYWSAPSDKEGYSGVGLLSRQCPLKVSYGIGEEEHDQEGRVIVAEFDSFVLVTAYVPNAGRGLVRLEYRQRWDEAFRKFLKGLAS
RKPLVLCGDLNVAHEEIDLRNPKGNKKNAGFTPQERQGFGELLQAVPLADSFRHLYPNTPYAYTFWTYMMNARSKNVGWRLDYFLLSHSLLPALCDSKIR
SKALGSDHCPITLYLAL
```

Fig. 3.6 Example of fasta format showing nucleotide base sequence of APEX nuclease 1 gene (*upper part*) and the corresponding protein sequence (*lower part*) of transcript ENST00000398030_D148E. The header line contains the gene, transcript or protein ID and description of the transcript is followed by a line containing the base sequence of the transcript. This gene contains a SNP of G→T at position 712 leading SAAV by replacing aspartic acid to glutamic acid at position 148 in the translated protein sequence. MS/MS spectra of peptide holding the SAAV and highlighted in bold in both sequences is shown in Fig. 3.11b. Non protein coding part is highlighted in *green*, the replaced D→E amino acid and GAG→GAT codon is highlighted in *red*, while the stop codon is shown in *blue* (highlights are only used to visualize different aspect of the sequence and is no part of the fasta format definition). In transcript T (thymine) is replaced by U (uracil)

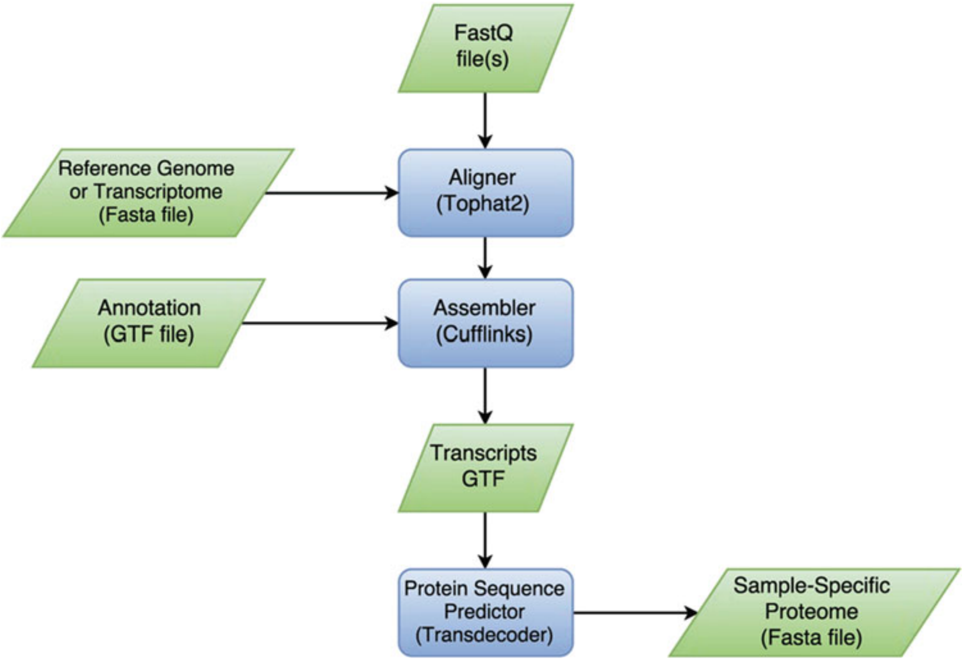

FastQ file(s)
Reference Genome or Transcriptome (Fasta file)
Aligner (Tophat2)
Annotation (GTF file)
Assembler (Cufflinks)
Transcripts GTF
Protein Sequence Predictor (Transdecoder)
Sample-Specific Proteome (Fasta file)

Fig. 3.7 Flow chart of bioinformatics workflow to pre-process sequencing data to make them ready for statistical analysis and provide the amino acid sequence of predicted translated proteins

mass spectrometry. Fragmenting and separating by LC is much more efficient for smaller peptides even though the enzymatic cleavage of proteins leads to much higher sample complexity.

The first problem related to shotgun LC-MS/MS proteomics is that the original sample protein composition with respect of protein species and quantities should be reconstructed from the primary amino acid sequences and quantities of the identified peptides. This operation is called protein inference (Farrah et al. 2011; Nesvizhskii 2007; Nesvizhskii and Aebersold 2005) and cannot be performed accurately because information on the intact protein species composition of the sample is lost during the enzymatic cleavage step. During protein identification, proteins that cannot be distinguished from each other based on the set of identified peptides are grouped in protein groups. Therefore, the quantity of a given protein in one group is always the same. Protein inference raises the question of how to determine the amount of single proteins included in the same group. Some methods only use the quantity of peptides that uniquely map to a protein group. Others split the quantities of shared peptides between protein groups according to the ratio of unique peptides, this fractional quantity of shared peptides is then used together with the complete quantities of unique peptides to calculate protein quantity. MaxQuant (Tyanova et al. 2015; Cox and Mann 2008) assigns the shared peptides (so called razor peptides) to a protein group with the largest number of identified peptides and uses the quantity of razor peptides in the assigned protein group to calculate the quantity of proteins present in that particular protein group. The many existing protein isoforms detected by RNA-Seq, which are included in the protein sequence Fasta file used for DBS, result in many identified proteins in protein groups as the outcome of a proteogenomics experiments. This outcome is better summarized as aggregate quantitative information of all protein products per gene, especially when only spectral counts are available, which only give semi-quantitative information. Further quantitative details should be explored at the peptide level, preferably using single-stage quantification, especially when single amino acid variants (SAAV) or short indels affect only one or two peptides of a target protein.

Peptide and protein quantification in comprehensive bottom-up LC-MS/MS experiments can be performed using stable isotope labelling and label-free approaches. Stable isotope labeling uses either metabolically incorporated stable isotopes such as the stable isotope labeling by amino acids in cell culture (SILAC) approach to incorporate ^{13}C and ^{15}N -labelled amino acids that cannot be synthesized *de novo* by cells in culture such as lysine and arginine, or the incorporation of ^{15}N into newly synthetized amino acids and thus into the complete newly synthetized proteome. Chemical labels may introduce moieties with different stable isotope constitutions that result in different MS signals either in single stage (*e.g.*, ICAT) or after fragmentation (*e.g.*, iTRAQ and TMT) for peptides originating from different samples. Stable isotope labeling techniques have the advantage of multiplexing, *i.e.*, reducing the number of analyses and instrument time by analyzing mixed samples, where sample specific information is obtained from ions with the same chemical but different isotopic constitution. This goes at the expense of the dynamic measurable concentration range according to the multiplexing factor. In label-free quantification the user has the choice between spectral count-based analyses based on counting the number of peptide-spectrum matches (PSMs) for each protein, which provide semi-quantitative peptide and protein quantification. The other option is to use the more accurate single-stage-MS-based quantification approach, which calculates the peak height, peak area or peak volume of isotopologue peaks in the single-stage MS map. For more information, the reader is advised to read specialized reviews on label-free (Christin et al. 2011; Horvatovich and Bischoff 2010; Horvatovich et al. 2006) and stable isotope-based quantification approaches (Bantscheff et al. 2007, 2012).

3.3.1 Raw Data

Mass spectrometry raw data is collected in scans, which is in nature one dimensional data with two parameters: m/z and ion intensity. However, the information content of scans depends on the applied mass spectrometry method. Nowadays, the untargeted comprehensive bottom-up data dependent acquisition (DDA) LC-MS/MS approach is the most commonly used approach. In DDA, a non-fragmented scan is first acquired that holds quantitative information on all compounds detected by the instrument at the time of the mass spectra acquisition. A single-stage scan is followed by 3 to 20 fragment ions scans obtained with a small precursor ion isolation window which is typically 1–2 Da wide and is centered to the most intense single-stage ions. The cycle containing single-stage scan and the 3–20 fragment ion scans with different precursor isolation windows is then repeated for the whole experiment, adopting dynamically to the actual peptide composition eluting from the LC column during the analysis, and results in fragmentation of the most abundant ions entering the mass spectrometer. The selected ions are then excluded with twice the peak width at half maximum to enable other lower abundant not yet fragmented peaks to be selected. Despite the m/z exclusion, DDA is biased towards high abundant peptides. The obtained fragment spectra (or MS/MS spectra) are then used for peptide identification, which means that MS/MS spectra are assigned to peptide primary amino acid sequences. Recently, data independent acquisition (DIA) (Sajic et al. 2015) is gaining popularity in which the precursor isolation window is larger, typically of 20–25 Da. The non-fragmented scan is followed by successive fragmented scans that have a precursor isolation window targeting different consecutive precursor m/z ranges. One full instrument duty cycle covers a large range of m/z ratios (typically between 300 and 2000 Da in proteomics applications) leading to 2–3 s of duty cycle. In theory, DIA data contains all the information that is possible to collect with an instrument that includes one stage fragmentation. DIA data is more complex and is more challenging to analyze and interpret than DDA fragment spectra obtained with small isolation windows that have low probability to have interferences i.e. fragment ions from multiple co-fragmented peptides.

The bioinformatics community is currently developing new solutions to analyze such data, such as OpenSWATH (Rost et al. 2014) or DIANA (Teleman et al. 2015). This chapter does not discuss the differences and properties of the different types of mass spectrometers and the reader is invited to visit reviews on this topics (Bensimon et al. 2012; Gstaiger and Aebersold 2009; Domon and Aebersold 2006).

The raw mass spectrometry data is generally saved by vendor data acquisition software in vendor specific binary formats, which are different from each other. To harmonize data storage, the HUPO protein standardization initiative (PSI) has established an xml based format for raw mass spectrometry data, such as mzXML (Pedrioli et al. 2004), mzData (Orchard et al. 2004) and mzML (Turewicz and Deutsch 2011), but older ASCII format such as Mascot Generic Format or mgf (Kirchner et al. 2010) are still used e.g. as input format for various data processing tools. Standardization of processed data for different purposes, such as to store peptide identification and protein inference results in mzIdentML format, to store quantification data in mzTab (Griss et al. 2014) and mzQuant (Walzer et al. 2013) formats and to exchange quality control metrics in qcML (Walzer et al. 2014) format have been developed by the proteomics community. The proteoWizard (Chambers et al. 2012; Kessner et al. 2008) toolset contains libraries and tools to convert raw vendor specific mass spectrometry data to HUPO PSI standard formats and enable the user to perform basic mass spectrometry signal processing operations. Raw mass spectrometry data can be visualized by multiple tools such as BatMass (Nesvizhskii and Avtonomov), TOPPView (Sturm and Kohlbacher 2009) from OpenMS (Bertsch et al. 2011) or PView (Khan et al. 2009).

3.3.2 Peptide Identification and Protein Inference

Primary peptide sequences are determined from fragment (MS/MS) spectra. The most widely used fragmentation approach is collision induced dissociation (CID), when ions of intact peptides are accelerated in a vacuum and collided with neutral gase. The collision is transferring energy to the peptides leading to cleavage of bonds in the peptide backbones Another type of fragmentation is electron transfer dissociation (ETD), which uses a negatively charged poly-aromatic compound such as fluoranthene, anthracene or azobenzene to transfer an electron to the positively charged peptide. The transferred electron conveys energy to the peptide backbone, which leads to fragmentation. There are three bonds that can lead to fragmentation on the peptide backbone leading to six types of fragments: a, b, c containing N-terminal and x, y and z containing the C-terminal of the peptide (Fig. 3.8a). However, not all fragments have the same probability to be observed in an MS/MS spectrum, for example, CID mainly leads to the formation of y ions, also resulting in lower abundance b ions and a ions can sometimes be observed. ETD fragmentation mainly leads to the formation of c and z ions. The ionization and fragmentation efficiency of intact peptides can be influenced by chemical modifications, for example by using chemical labels that contain basic residues or residues that can provide a mobile proton (Bischoff et al. 2015). In the fragmentation process, the lower energy bonds will be cleaved, which often results in an incomplete fragment ion series (Fig. 3.8b). This prevents the *de novo* interpretation of the mass spectra, which prevent identification of the MS/MS spectra if the user does not have any presumptions on the peptides sequence. Additional fragment mass spectra may contain considerable noise, which further complicates the identification process. For this reason, the best approach to interpret such data is to use a list of protein sequences that are supposed to be present in the analyzed samples. Such protein sequence can be predicted from the genome of the host organism, which contains the most prevalent protein sequences or the so-called canonical sequences. UniProt (Consortium 2015) and Ensembl (Herrero et al. 2016) provide high quality canonical sequences that are used for peptide and protein identification during normal proteomics data analysis. One must note that the canonical

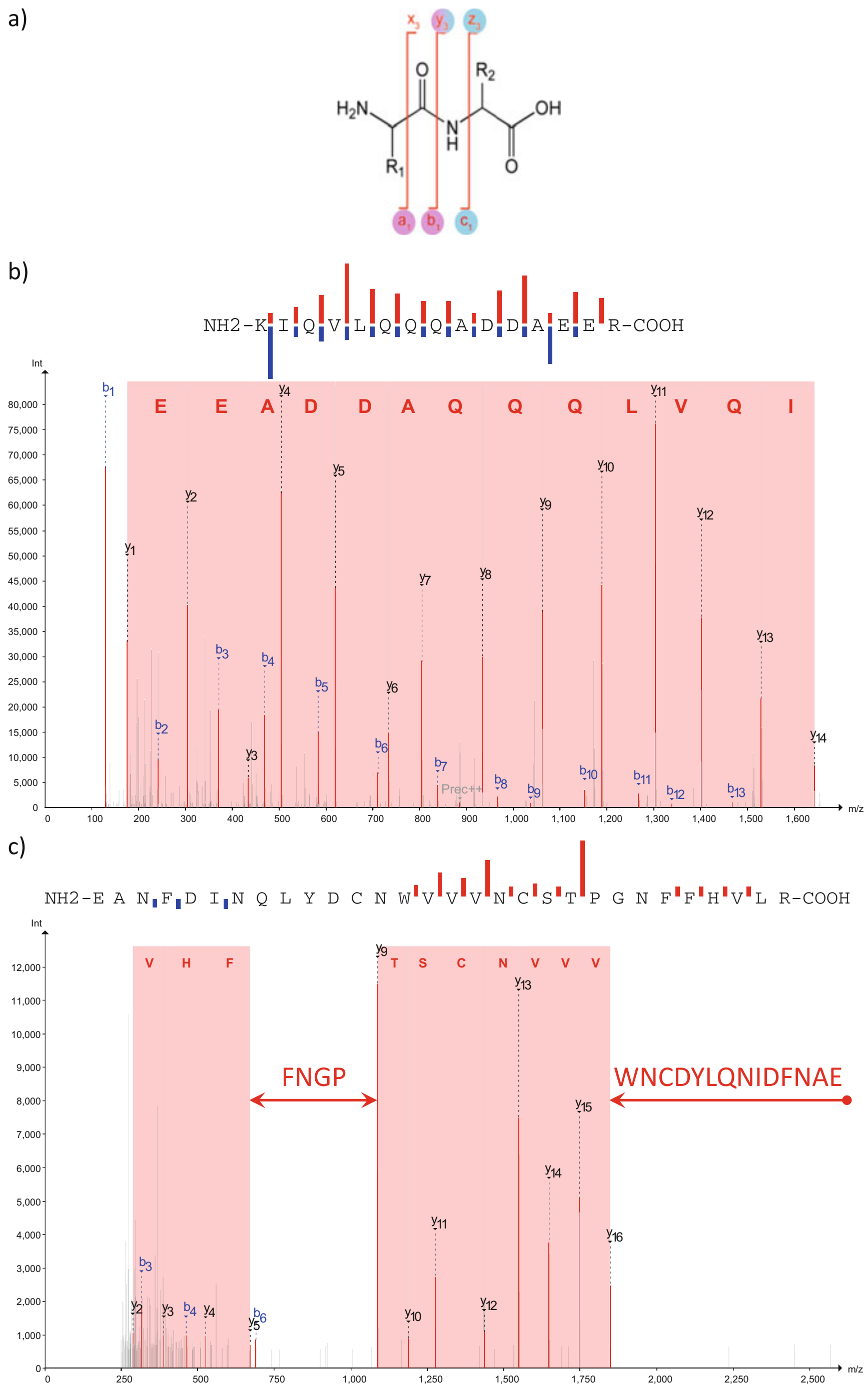


Fig. 3.8 (**a**) schematic representation of fragment ion series (a, b, c for N-terminal and x, y, z for C terminal) generated during peptide backbone fragmentation. y, b and a ions are generated mainly during CID (*purple*) fragmentation, while c, z and with lower abundance y ions are generated during ETD (*blue*) fragmentation. (**b**) CID MS/MS spectra of KIQVLQQQADDAEER peptide showing complete y and almost complete b ion series, which spectra is suitable for *de novo* interpretation. (**c**) MS/MS of EANFDINQLYDCNWVVVVNCSTPGNFFHVLR peptides, which shows incomplete y and b ion series and gaps in y ion series are highlighted with *red arrow* indicating the missing sequence part. These gaps prevent *de novo* interpretation since the exact amino acid sequence information is missing. In MS/MS spectra non-identified signals are highlighted in *grey*. These signals may correspond to noise, non-interpreted fragment ions or fragment ions from co-eluting peptides that fell into the precursor ion selection window. Visualisation made with PeptideShaker (Vaudel et al. 2015) and the figure is adapted with permission from Bischoff et al. (2015) (Copyright (2015) Elsevier)

sequence contains the sequence of most prevalent protein form, which is the most similar orthologue sequence to other species and the length of the protein form that allow the clearest description of the protein sequence variability (see Introduction Sect. 3.1). This protein sequence does not allow identification of all protein sequence variants, especially those that are specific to individuals and may bear importance in disease mechanisms.

In an LC-MS/MS dataset not all MS/MS spectra are identified during DBS, due to the following reasons:

1. The fragment spectra is too noisy
2. The fragmentation efficiency is too low to perform accurate identification
3. The absence of the peptide sequence in the protein sequence database
4. The presence of PTMs not searched during DBS

Sharing raw LC-MS/MS data and reusing it by several bioinformatics portals *e.g.*, to catalogue identified peptides and provide high quality spectral libraries such as PeptideAtlas (Deutsch et al. 2008; Farrah et al. 2011) is promoted by ProteomeXchange (Ternent et al. 2014; Cote et al. 2012), which is an initiative of the European Bioinformatics Institute to store raw proteomics mass spectrometry data.

Due to gaps in fragment ion series and noise in fragment spectra, the most successful strategy is DBS. In this process the sequence of proteins supposed to be present in the sample are digested *in silico* with the protease used for the protein cleavage in the experiment and peptides that have the same theoretical mass (with certain mass tolerance) than the precursor ion are selected. The mass of high abundant ion series of the selected peptides are *in silico* calculated and the obtained mass list is compared with the mass list of the MS/MS spectra using score specific to the DBS algorithm. The peptide with the highest score if it pass the threshold with given false discovery rate (FDR) is then considered to be the identity of the MS/MS spectra (Fig. 3.9). Scores are generally dependent from multiple parameters, such as the size of the search space, *i.e.*, how well does the protein sequence database match the measured proteome (Shanmugam and Nesvizhskii 2015), the considered PTMs of peptides, the mass resolution of the precursor and fragment ions, and the fragmentation efficiency and quality (noise content) of the MS/MS spectra. Additionally, not all MS/MS spectra will have a corresponding match in the search space; such spectra will be matched and scored erroneously. For this reason, the goal is to find scores that can separate correct identifications from the incorrect ones with well described statistics such as false discovery rate (FDR).

The score distribution of correct and incorrect identifications should be determined to calculate FDR, and there are two main widely used approaches:

1. Expectation-Maximization (EM) (Keller et al. 2002) approach based on empirical Bayesian statistics
2. Target-decoy approach (TD) (Elias and Gygi 2010)

EM tries to identify the score distribution of the correct and incorrect hits by calculating two distinct distributions based on the mixture model. While the TD approach tries to determine the distribution of the incorrect identifications based on decoy peptide sequences generally obtained by *in silico* digestion of reversed protein sequences used for the DBS. These approaches allow the user to obtain a list of identified PSMs, which can be used to derive a list of identified unique peptides, which can then be used to perform protein inference.

Since peptides are measured and identified in shotgun LC-MS/MS experiments, the original protein constitution of the samples should be reconstructed based on the identified set of peptides (Fig. 3.10). This is not a trivial task, since identified peptides sequence may map uniquely to a protein sequence or be shared between multiple ones. The other difference between sequencing and proteomics data is the scale of the number

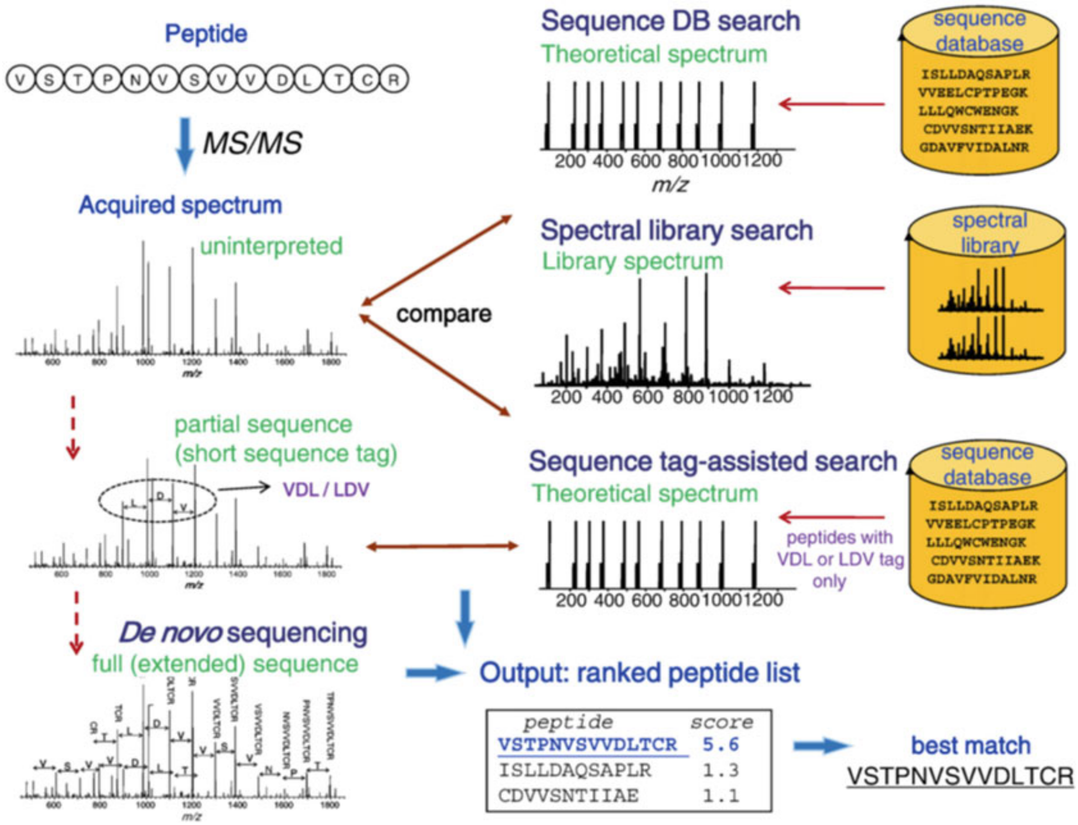


Fig. 3.9 Schematic representation of bioinformatics algorithms performing peptide spectrum matches (*PSM*). Acquired raw MS/MS spectra are either submitted to (1). DBS that match list of fragments ions predicted from sequence that supposed to be present in the sample, (2). submitted to spectral library search, which match the raw MS/MS to a library of annotated MS/MS spectra, (3). submitted to sequence tag search (or mass-tag) algorithm or (4). *de novo* sequencing. The output of the search is a ranked list of peptides, where the peptide sequence with best score is considered as best match and peptide sequence of the analyzed MS/MS spectra. The scores of the best matches are submitted to FDR calculation either using empirical expectation-maximization algorithm or target-decoy approach (Figure adapted with permission from (Nesvizhskii 2010). Copyright (2010) Elsevier)

of entries. The number of identified peptide sequences is much lower (typically 10,000–30,000 unique sequences) than the number of uniquely mapping reads (typically 20 millions reads). The overlap between the peptide sequences is low, which generally occurs between peptides having missed cleavages (locations in the protein where the enzyme should cut in theory, but did not cut to produce a peptide).

Identified proteins are grouped together when they cannot be distinguished from each other based on the set of observed peptide sequences in the dataset. The sequence coverage of the identified protein is an important parameter. The sequence coverage depends on the abundance of the protein and the peptide composition. The most abundant proteins have higher sequence coverage than lower abundant proteins. The average protein sequence coverage is low, with a medium of 10–20 % in a typical proteomics dataset. This means that peptides that could distinguish various protein isoforms, due to for instance splice junction differences, SAAV or small indels, are incomplete even when deep sequenc-

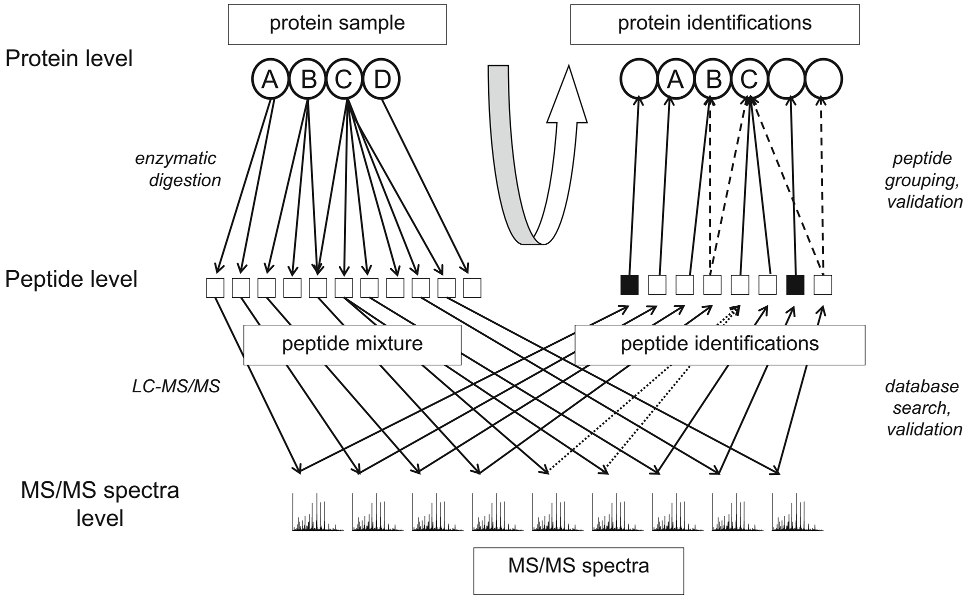


Fig. 3.10 Schematic representation of shotgun LC-MS/MS analysis of a proteomics samples and followed bioinformatics data interpretation. Original protein constitution of a samples is disrupted by enzymatic cleavage resulting in a highly complex peptide mixture analyzed by LC-MS/MS. The obtained MS/MS of peptides are then identified with DBS or by other tools and the set if highly confidently identified peptides are used to construct back the original protein constitution of the sample by performing protein inference. *Black squares* in peptide identifications represent wrong PSM, which lead to include incorrectly identified peptides and proteins (Figure adapted with permission from Nesvizhskii et al. (2003) Copyright (2003) American Chemistry Society)

ing is performed (Ruggles et al. 2015; Tay et al. 2015; Sheynkman et al. 2013). The low sequence coverage is caused by multiple factors:

1. Proteins and peptide signals cannot be amplified (as is the case with DNA and RNA signals)
2. Not all MS/MS fragment spectra are identified
3. The applied protease (*e.g.*, trypsin) does not provide unique or enough protein sequence specific peptides for the complete sequence of the analyzed proteins

Sequence coverage can be improved by deep sequencing that use multilevel fractionation, *e.g.*, by applying multidimensional chromatographic separation (Horvatovich et al. 2010), by using different peptide fragmentation approaches in the mass spectrometer and chemical labels that enhance fragmentation efficiency (Bischoff et al. 2015) and by using multiple proteases for enzymatic cleavage (Low et al. 2013; Trevisiol et al. 2015).

The false identification of peptides may lead to the incorrect identification of a protein and the fact that multiple correctly identified peptides map to a single protein, while incorrectly identified peptides map randomly to single proteins in the database leads to an enrichment of false protein identifications compared to PSM or peptide identification errors. For this reason, the FDR rate should be calculated not only for the PSM and peptide but also at the protein level (Vaudel et al. 2015).

Beside DBS, other approaches can be used to perform PSM. The short sequence tag approach tries to identify consecutive amino acid sequence in the MS/MS spectra and uses the precursor ion mass and the masses of the fragments from the N, and the C terminus of peptides for the identification. MS/MS spectra that include low noise content and shows complete fragment ion series could be used for hypothesis-free *de novo* sequencing, without the use of any assumptions on protein sequence that should be present in the analyzed sample. A more and more popular approach is the use of spectral similarities between the MS/MS spectra of interest and high quality identified MS/MS spectra (so called consensus spectra) often averaged from multiple MS/MS spectra of different experiments. This approach is called spectral library search (Lam 2011) and has the advantage that it does not only use the mass list of the fragment ions, but that it also includes their intensity, which is a parameter that is difficult to predict *in silico* and which is not or only partially included in DBS. More and more high quality peptide spectral libraries are available that can be used to perform spectral library searches. High quality annotated spectral libraries are available such as NIST, the PeptideAtlas (Deutsch et al. 2008) and the Global Proteome Machine Database (GPMDB) (Craig et al. 2006). Figure 3.9 provides a summary of the most important PSM identification strategies.

In proteogenomics, the FDR rate of MS/MS identifications of novel peptide differs from peptides derived from canonical sequences of public databases, such as UniProt (Consortium 2015). For this reason the best PSM scoring strategy is cascade identification, which includes consecutive steps of identification as follow:

1. Filter out all low quality MS/MS spectra
2. DBS identification using UniProt database (SwissProt and TrEMBL) or Ensembl
3. Identification of the remaining non-identified MS/MS spectra with novel peptide or protein sequences (Nesvizhskii 2014)

Similar cascade identification strategies have been implemented for different types of rare peptides, such as non- and semi-tryptic peptides, terminal peptides and PTM searches as described in Kertesz-Farkas *et al.* (Kertesz-Farkas et al. 2015).

Many software tools exist to perform PSM, protein inference using a given set of FDR at PSM, peptide and protein levels, these include the Trans Proteomic Pipeline (Deutsch et al. 2010; Deutsch et al. 2015) (TPP, open source), the TOPPAS workflow, which is based on OpenMS for label-free quantification and identification (Weisser et al. 2013), MaxQuant (Cox and Mann 2008) (open source), SearchGUI (Vaudel et al. 2011) / PeptideShaker (Vaudel et al. 2015) (open source) and PEAKS (commercial) (Zhang et al. 2012). Many individual tools exist for DBS (Eng et al. 2013; Kim and Pevzner 2014; Bjornson et al. 2008; Geer et al. 2004), *de novo* sequencing (Muth et al. 2014; Jeong et al. 2013; Frank and Pevzner 2005) and FDR calculations at PSM, at peptide and protein levels (Kall et al. 2007). For further details on these tools, the reader is invited to read specialized reviews on the topic (Hoopmann and Moritz 2013; Eng et al. 2011; Hughes et al. 2010; Kapp and Schutz 2007).

3.4 Applications, Conclusion and Future Perspectives

Acquiring genomics (mainly polyadenylated mRNA) and shotgun proteomics data from the same sample and evaluate it in a proteogenomics data integration pipeline allows to gain information at both molecular levels but also to identify novel protein forms that would not be identified using public databases with DBS. As an example, we present the data of the proteogenomics analysis of the human lung fibroblast cell line MRC5. Using the standard identification of UniProt we identified 11,936 peptides and when we used the RNA sequence information of the same cells we could identify an additional 282 peptides, which represent the sample specific peptide sequence. Figure 3.11a shows a number of peptide sequences that has been identified with canonical sequences of UniProt, peptides that match to SAAVs due to non-synonymous SNPs, peptides matching to new isoforms and peptides that match to non-annotated new gene models. Figure 3.11b shows an example of a high quality MS/MS spectrum presented with complete y and b ion annotation of peptides (VSYGIG(D→E) EEHDQEGR) holding SAAV that replaces an aspartic acid (D) to glutamic acid (E) at positon 148. This peptide is mapping uniquely to APEX

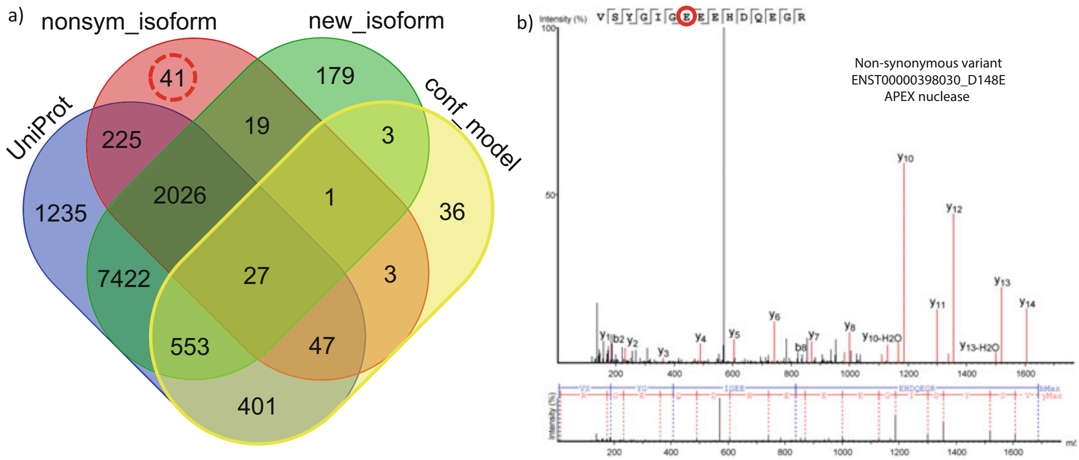


Fig. 3.11 Identification of peptides that match uniquely to sample specific protein forms in human MRC5 fibroblast cell line. (**a**) Venn diagram representing the number of peptides that has been identified using protein sequence database containing sequence from UniProt, non-synonymous isoforms, new isoforms and new gene models. (**b**) example of CID MS/MS spectra of VSYGIG(D → E)EEHDQEGR with annotation of y and b ion series of peptides that hold SAAV replacing aspartic acid (D) to Glutamic acid (E) at position 148. Base sequence of the corresponding gene and the amino acid sequence of the corresponding protein is shown in Fig. 3.6, highlighting the gene structure, the presented peptide sequence and the position of SAAV

nuclease, which is a multifunctional DNA repair enzyme. This peptide cannot be identified in the human UniProt protein sequence, but can be found in the APEX nuclease sequence of many other species, which indicates that this mutation may alter the activity of this protein in the MRC5 human cell line.

The moderate spearman correlation of 0.4 between the amount of transcript coding proteins and proteins shows that there is an additional level of regulation which includes post-transcriptional and post-translational effects (Schwanhausser et al. 2013, 2011). Therefore, the information at the two molecular levels differs and should be considered to be complementary. Both levels may deliver large amount of information, which is difficult to interpret, such as number of differentially expressed proteins. In this case, focusing on the intersection of genes and transcripts / proteins that show the same trend at both molecular levels may provide a useful focus to interpret the outcome of a proteogenomics study.

An example of considering joint changes at the transcriptomics and proteomics levels is shown in Fig. 3.12. This figure shows a pseudo Volcano plot indicating a fold change and *t*-test significance at transcript and protein levels. The result was obtained in a proteogenomics study performed to identify molecular changes in liver of hypertensive SHR rats when compared to a control BN-L*x* rat strain. The study by Low et al. (2013) shows genome and polyadenylated transcriptome sequencing for eight rats (about 100 million of reads/sample) and deep proteomics analysis for two rats using a two dimensional LC-MS/MS experiment. To obtain the highest possible sequence coverage and the largest measured dynamic concentration range, five proteases (trypsin, chymotrypsin, LysC, GluC and AspN) and strong cation exchange (SCX) as first liquid chromatography and reversed phase C18 (RPC18) with low pH as second dimension were used. This setup led to 36 fractions / samples and 180 RPC18 analysis using high resolution Orbitrap instrument and nearly 2 weeks of analysis time. From the acquired 12 million MS/MS spectra, two million were identified using Mascot / PEAKS DBS searches and resulted in 175,000 non-redundant peptide sequences matching to 26,463 rat proteins. In this experiment, 1195 predicted new genes, 83 splicing

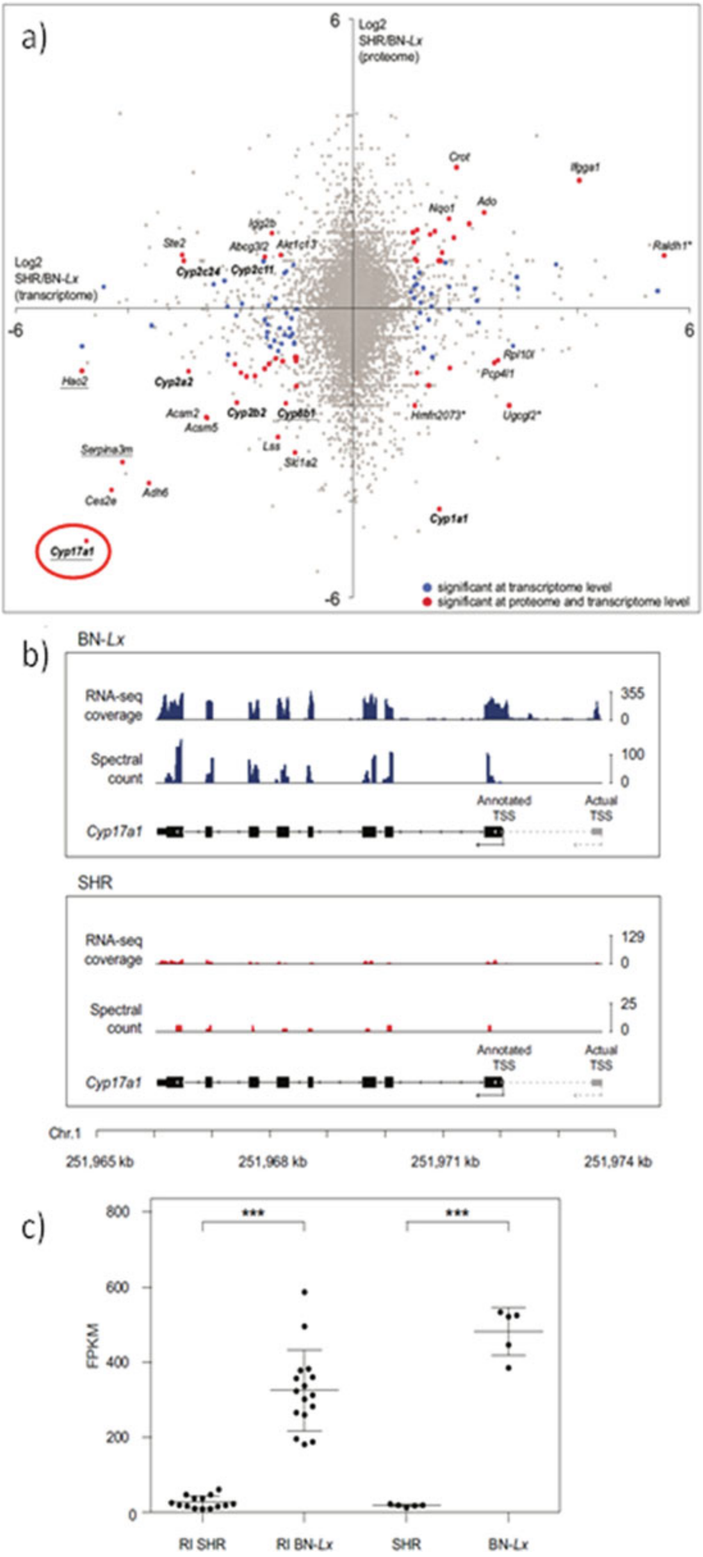


Fig. 3.12 Outcome of proteogenomics study in hypertensive SHR and control BN-L*x* rats. (**a**) pseudo Volcano plot showing fold changes in transcript on the horizontal axis and fold changes of proteins in the vertical axis. *Blue dots* (n = 59) represent significant changes at transcriptome level only and *red dots* (n = 54) represent significant changes at transcript and protein levels. The most significantly down-regulated Cyp17a1 gene is highlighted with red circle. (**b**) transcript and protein expression level at location of gene Cyp17a1 showing the position of incorrectly annotated start site (TSS *black arrow*) and the real start site (TSS *grey arrow*). (**c**) expression quantitative trait loci (eQTL) showing the transcript expression regulation by the SNP at the real start site (Adapted with permission from Low et al. (2013) Copyright (2013) Elsevier)

events, 126 proteins with non-synonymous variants and 20 isoforms with non-synonymous RNA editing were identified.

Differential gene expression analysis at both molecular layers revealed that genes related to cytochrome P450 (CYP450) are mainly differentially expressed in the same direction. Particularly, the gene Cyp17a1 was the strongest down-regulated in hypertensive SHR rats. Having both genomics and transcriptomics data in hand, it was demonstrated that the transcription start site was incorrectly annotated in the reference rat genome and that the correct start site was 2 kb further upstream from the current annotation on the 5’ exon. The correct start site in SHR rats included a SNP, which prevented transcription and translation of the protein coded by the Cyp17a1 gene (Fig. 3.11b). In this case the proteogenomics analysis helped to identify a gene related to the hypertensive rat phenotype, but also to correct the genome annotation, revealing the cause of the down-regulation of the transcript and the protein product by a SNP at the starting site of the Cyp17a1 gene.

Proteogenomics still requires important efforts to collect data at genomic and/or transcriptomic and proteomics levels and the correct analysis of the obtained data, which requires expertise from both omics fields as well as from bioinformatics. Despite the significant improvement of high-throughput proteomics peptide identification technology in recent years, proteomics still does not provide clean data for *de novo* sequencing and is unable to deliver the same coverage of peptide sequence information when compared to genomics sequencing technology. Additional improvement will be possible by combining ribosomal sequencing data, the so-called translatome, with transcriptomics, since it helps to filter out transcripts that have a low potential for translation and may include potentially translated lncRNA – despite the fact that this technology delivers only 30 nucleotide base length sequences (Gawron et al. 2014; Chang et al. 2014). Proteogenomics analysis can be completed with the PUromycin-associated Nascent CHain Proteomics (PUNCH-P) technology that aims to identify newly synthetized proteins by capturing ribosome-nascent chain complexes from cells followed by incorporation of biotinylated puromycin (Aviner et al. 2013).

Further impetus for proteogenomics is evident in the Chromosome-Centric Human Proteome Project (C-HPP) (Horvatovich et al. 2015), which aims to catalogue all human protein products and make them searchable on the basis of genomics location. Proteogenomics data acquisition and data integration plays a central role in C-HPP, which promotes the development of new technologies and bioinformatics workflows with strong quality control, and aims to provide a powerful technology platform for clinical application and personalized medicine.

References

Aviner, R., Geiger, T., & Elroy-Stein, O. (2013). PUNCH-P for global translatome profiling: Methodology, insights and comparison to other techniques. *Translation* (Austin), *1*(2), e27516. doi:10.4161/trla.27516

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry, 389*(4), 1017–1031. doi:10.1007/s00216-007-1486-6.

Bantscheff, M., Lemeer, S., Savitski, M. M., & Kuster, B. (2012). Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry, 404*(4), 939–965. doi:10.1007/s00216-012-6203-4.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Research, 41*(Database issue), D991–D995. doi:10.1093/nar/gks1193.

Bensimon, A., Heck, A. J., & Aebersold, R. (2012). Mass spectrometry-based proteomics and network biology. *Annual Review of Biochemistry, 81*, 379–405. doi:10.1146/annurev-biochem-072909-100424.

Bertsch, A., Gropl, C., Reinert, K., & Kohlbacher, O. (2011). OpenMS and TOPP: Open source software for LC-MS data analysis. *Methods in Molecular Biology, 696*, 353–367. doi:10.1007/978-1-60761-987-1_23.

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research, 29*(12), 2607–2618.

Bischoff, R., & Schlüter, H. (2012). Amino acids: Chemistry, functionality and selected non-enzymatic post-translational modifications. *Journal of Proteomics, 75*(8), 2275–2296. doi:10.1016/j.jprot.2012.01.041.

Bischoff, R., Permentier, H., Guryev, V., & Horvatovich, P. (2015). Genomic variability and protein species – Improving sequence coverage for proteogenomics. *Journal of Proteomics*. doi:10.1016/j.jprot.2015.09.021.

Bjornson, R. D., Carriero, N. J., Colangelo, C., Shifman, M., Cheung, K. H., Miller, P. L., & Williams, K. (2008). X!!Tandem, an improved method for running X! Tandem in parallel on collections of commodity computers. *Journal of Proteome Research, 7*(1), 293–299. doi:10.1021/pr0701198.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170.

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., & Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology, 30*(10), 918–920. doi:10.1038/nbt.2377.

Chang, C., Li, L., Zhang, C., Wu, S., Guo, K., Zi, J., Chen, Z., Jiang, J., Ma, J., Yu, Q., Fan, F., Qin, P., Han, M., Su, N., Chen, T., Wang, K., Zhai, L., Zhang, T., Ying, W., Xu, Z., Zhang, Y., Liu, Y., Liu, X., Zhong, F., Shen, H., Wang, Q., Hou, G., Zhao, H., Li, G., Liu, S., Gu, W., Wang, G., Wang, T., Zhang, G., Qian, X., Li, N., He, Q. Y., Lin, L., Yang, P., Zhu, Y., He, F., & Xu, P. (2014). Systematic analyses of the transcriptome, translatome, and proteome provide a global view and potential strategy for the C-HPP. *Journal of Proteome Research, 13*(1), 38–49. doi:10.1021/pr4009018.

Christin, C., Bischoff, R., & Horvatovich, P. (2011). Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC-MS for biomarker discovery. *Talanta, 83*(4), 1209–1224. doi:10.1016/j.talanta.2010.10.029.

Chuh, K. N., & Pratt, M. R. (2015). Chemical methods for the proteome-wide identification of posttranslationally modified proteins. *Current Opinion in Chemical Biology, 24*, 27–37. doi:10.1016/j.cbpa.2014.10.020.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research, 38*(6), 1767–1771. doi:10.1093/nar/gkp1137/ConsortiumN.

Consortium U. (2015). UniProt: A hub for protein information. *Nucleic Acids Research, 43*(Database issue), D204–D212. doi:10.1093/nar/gku989.

Cote, R. G., Griss, J., Dianes, J. A., Wang, R., Wright, J. C., van den Toorn, H. W., van Breukelen, B., Heck, A. J., Hulstaert, N., Martens, L., Reisinger, F., Csordas, A., Ovelleiro, D., Perez-Rivevol, Y., Barsnes, H., Hermjakob, H., & Vizcaino, J. A. (2012). The PRoteomics IDEntification (PRIDE) Converter 2 framework: An improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Molecular & Cellular Proteomics, 11*(12), 1682–1689. doi:10.1074/mcp.O112.021543.

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology, 26*(12), 1367–1372. doi:10.1038/nbt.1511.

Craig, R., Cortens, J. C., Fenyo, D., & Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research, 5*(8), 1843–1849. doi:10.1021/pr0602085.

Deutsch, E. W., Lam, H., & Aebersold, R. (2008). PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. *EMBO Reports, 9*(5), 429–434. doi:10.1038/embor.2008.56.

Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., & Aebersold, R. (2010). A guided tour of the trans-proteomic pipeline. *Proteomics, 10*(6), 1150–1159. doi:10.1002/pmic.200900375.

Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., & Moritz, R. L. (2015). Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clinical Applications, 9*(7–8), 745–754. doi:10.1002/prca.201400164.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15–21. doi:10.1093/bioinformatics/bts635.

Domon, B., & Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science, 312*(5771), 212–217. doi:10.1126/science.1124619.

Elias, J. E., & Gygi, S. P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in Molecular Biology, 604*, 55–71. doi:10.1007/978-1-60761-444-9_5.

Eng, J. K., Searle, B. C., Clauser, K. R., & Tabb, D. L. (2011). A face in the crowd: Recognizing peptides through database search. *Molecular & Cellular Proteomics, 10*(11), R111.009522. doi:10.1074/mcp.R111.009522.

Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *Proteomics, 13*(1), 22–24. doi:10.1002/pmic.201200439.

Farrah, T., Deutsch, E. W., Omenn, G. S., Campbell, D. S., Sun, Z., Bletz, J. A., Mallick, P., Katz, J. E., Malmstrom, J., Ossola, R., Watts, J. D., Lin, B., Zhang, H., Moritz, R. L., & Aebersold, R. (2011). A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Molecular & Cellular Proteomics, 10*(9), M110 006353. doi:10.1074/mcp.M110.006353.

Fiume, M., Williams, V., Brook, A., & Brudno, M. (2010). Savant: Genome browser for high-throughput sequencing data. *Bioinformatics, 26*(16), 1938–1944. doi:10.1093/bioinformatics/btq332.

Frank, A., & Pevzner, P. (2005). PepNovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry, 77*(4), 964–973.

Gawron, D., Gevaert, K., & Van Damme, P. (2014). The proteome under translational control. *Proteomics, 14*(23–24), 2647–2662. doi:10.1002/pmic.201400165.

Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., & Bryant, S. H. (2004). Open mass spectrometry search algorithm. *Journal of Proteome Research, 3*(5), 958–964. doi:10.1021/pr0499491.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology, 29*(7), 644–652. doi:10.1038/nbt.1883.

Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G. G., Salek, R. M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q. W., Del Toro, N., Perez-Riverol, Y., Ghali, F., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaino, J. A., & Hermjakob, H. (2014). The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics, 13*(10), 2765–2775. doi:10.1074/mcp.O113.036681.

Gstaiger, M., & Aebersold, R. (2009). Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics, 10*(9), 617–627. doi:10.1038/nrg2633.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (2016). Ensembl comparative genomics resources. *Database: The Journal of Biological Databases and Curation*. doi:10.1093/database/bav096.

Hoopmann, M. R., & Moritz, R. L. (2013). Current algorithmic solutions for peptide-based proteomics data generation and identification. *Current Opinion in Biotechnology, 24*(1), 31–38. doi:10.1016/j.copbio.2012.10.013.

Horvatovich, P. L., & Bischoff, R. (2010). Current technological challenges in biomarker discovery and validation. *European Journal of Mass Spectrometry, 16*(1), 101–121. doi:10.1255/ejms.1050.

Horvatovich, P., Govorukhina, N., & Bischoff, R. (2006). Biomarker discovery by proteomics: Challenges not only for the analytical chemist. *The Analyst, 131*(11), 1193–1196. doi:10.1039/b607833h.

Horvatovich, P., Hoekman, B., Govorukhina, N., & Bischoff, R. (2010). Multidimensional chromatography coupled to mass spectrometry in analysing complex proteomics samples. *Journal of Separation Science, 33*(10), 1421–1437. doi:10.1002/jssc.201000050.

Horvatovich, P., Lundberg, E. K., Chen, Y. J., Sung, T. Y., He, F., Nice, E. C., Goode, R. J., Yu, S., Ranganathan, S., Baker, M. S., Domont, G. B., Velasquez, E., Li, D., Liu, S., Wang, Q., He, Q. Y., Menon, R., Guan, Y., Corrales, F. J., Segura, V., Casal, J. I., Pascual-Montano, A., Albar, J. P., Fuentes, M., Gonzalez-Gonzalez, M., Diez, P., Ibarrola, N., Degano, R. M., Mohammed, Y., Borchers, C. H., Urbani, A., Soggiu, A., Yamamoto, T., Salekdeh, G. H., Archakov, A., Ponomarenko, E., Lisitsa, A., Lichti, C. F., Mostovenko, E., Kroes, R. A., Rezeli, M., Vegvari, A., Fehniger, T. E., Bischoff, R., Vizcaino, J. A., Deutsch, E. W., Lane, L., Nilsson, C. L., Marko-Varga, G., Omenn, G. S., Jeong, S. K., Lim, J. S., Paik, Y. K., & Hancock, W. S. (2015). Quest for missing proteins: Update 2015 on chromosome-centric human proteome project. *Journal of Proteome Research, 14*(9), 3415–3431. doi:10.1021/pr5013009.

Hughes, C., Ma, B., & Lajoie, G. A. (2010). De novo sequencing methods in proteomics. *Methods in Molecular Biology, 604*, 105–121. doi:10.1007/978-1-60761-444-9_8.

Jeong, K., Kim, S., & Pevzner, P. A. (2013). UniNovo: A universal tool for de novo peptide sequencing. *Bioinformatics, 29*(16), 1953–1962. doi:10.1093/bioinformatics/btt338.

Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods, 4*(11), 923–925. doi:10.1038/nmeth1113.

Kapp, E., & Schutz, F. (2007). Overview of tandem mass spectrometry (MS/MS) database search algorithms. Current protocols in protein science / editorial board, John E Coligan [et al] Chapter 25:Unit25 22. doi:10.1002/0471140864.ps2502s49.

Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry, 74*(20), 5383–5392.

Kertesz-Farkas, A., Keich, U., & Noble, W. S. (2015). Tandem mass spectrum identification via cascaded search. *Journal of Proteome Research, 14*(8), 3027–3038. doi:10.1021/pr501173s.

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics, 24*(21), 2534–2536. doi:10.1093/bioinformatics/btn323.

Khan, Z., Bloom, J. S., Garcia, B. A., Singh, M., & Kruglyak, L. (2009). Protein quantification across hundreds of experimental conditions. *Proceedings of the National Academy of Sciences of the United States of America, 106*(37), 15544–15548. doi:10.1073/pnas.0904100106.

Kim, S., & Pevzner, P. A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications, 5*, 5277. doi:10.1038/ncomms6277.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*(4), R36. doi:10.1186/gb-2013-14-4-r36.

Kirchner, M., Steen, J. A., Hamprecht, F. A., & Steen, H. (2010). MGFp: An open Mascot Generic Format parser library implementation. *Journal of Proteome Research, 9*(5), 2762–2763. doi:10.1021/pr100118f.

Lam, H. (2011). Building and searching tandem mass spectral libraries for peptide identification. *Molecular & Cellular Proteomics, 10*(12), R111.008565. doi:10.1074/mcp.R111.008565.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., 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, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., 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, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., 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, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., 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, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., & Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860–921. doi:10.1038/35057062.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352.

Low, T. Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hubner, N., van Breukelen, B., Mohammed, S., Cuppen, E., Heck, A. J., & Guryev, V. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Reports, 5*(5), 1469–1478. doi:10.1016/j.celrep.2013.10.041.

Markiv, A., Rambaruth, N. D., & Dwek, M. V. (2012). Beyond the genome and proteome: Targeting protein modifications in cancer. *Current Opinion in Pharmacology, 12*(4), 408–413. doi:10.1016/j.coph.2012.04.003.

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics, 12*(10), 671–682. doi:10.1038/nrg3068.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297–1303. doi:10.1101/gr.107524.110.

Menschaert, G., & Fenyo, D. (2015). Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrometry Reviews*. doi:10.1002/mas.21483.

Metzker, M. L. (2010). Sequencing technologies – The next generation. *Nature Reviews Genetics, 11*(1), 31–46. doi:10.1038/nrg2626.

Muth, T., Weilnbock, L., Rapp, E., Huber, C. G., Martens, L., Vaudel, M., & Barsnes, H. (2014). DeNovoGUI: An open source graphical user interface for de novo sequencing of tandem mass spectra. *Journal of Proteome Research, 13*(2), 1143–1146. doi:10.1021/pr4008078.

Nesvizhskii, A. I. (2007). Protein identification by tandem mass spectrometry and sequence database searching. *Methods in Molecular Biology, 367*, 87–119. doi:10.1385/1-59745-275-0:87.

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics, 73*(11), 2092–2123. doi:10.1016/j.jprot.2010.08.009.

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods, 11*(11), 1114–1125. doi:10.1038/nmeth.3144.

Nesvizhskii, A., & Avtonomov, D. http://www.batmass.org/

Nesvizhskii, A. I., & Aebersold, R. (2005). Interpretation of shotgun proteomic data: The protein inference problem. *Molecular & Cellular Proteomics, 4*(10), 1419–1440. doi:10.1074/mcp.R500012-MCP200.

Nesvizhskii, A. I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry, 75*(17), 4646–4658.

Orchard, S., Taylor, C., Hermjakob, H., Zhu, W., Julian, R., & Apweiler, R. (2004). Current status of proteomic standards development. *Expert Review of Proteomics, 1*(2), 179–183. doi:10.1586/14789450.1.2.179.

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PloS One, 7*(2), e30619. doi:10.1371/journal.pone.0030619.

Pearson, W. R., Wood, T., Zhang, Z., & Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics, 46*(1), 24–36. doi:10.1006/geno.1997.4995.

Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., & Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology, 22*(11), 1459–1466. doi:10.1038/nbt1031.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology, 29*(1), 24–26. doi:10.1038/nbt.1754.

Rost, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmstrom, J., Malmstrom, L., & Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology, 32*(3), 219–223. doi:10.1038/nbt.2841.

Ruggles, K. V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M. D., Clauser, K. R., Tabb, D. L., Mertins, P., Slebos, R., Erdmann-Gilmore, P., Li, S., Gunawardena, H. P., Xie, L., Liu, T., Zhou, J. Y., Sun, S., Hoadley, K. A., Perou, C. M., Chen, X., Davies, S. R., Maher, C. A., Kinsinger, C. R., Rodland, K. D., Zhang, H., Zhang, Z., Ding, L., Townsend, R. R., Rodriguez, H., Chan, D., Smith, R. D., Liebler, D. C., Carr, S. A., Payne, S., Ellis, M. J., & Fenyo, D. (2015). An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Molecular & Cellular Proteomics*. doi:10.1074/mcp.M115.056226.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., & Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *eLife, 3*, e03523. doi:10.7554/eLife.03523.

Sajic, T., Liu, Y., & Aebersold, R. (2015). Using data-independent, high-resolution mass spectrometry in protein biomarker research: Perspectives and clinical applications. *Proteomics Clinical Applications, 9*(3–4), 307–321. doi:10.1002/prca.201400117.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature, 265*(5596), 687–695.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature, 473*(7347), 337–342. doi:10.1038/nature10098.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M. (2013). Corrigendum: Global quantification of mammalian gene expression control. *Nature, 495*(7439), 126–127. doi:10.1038/nature11848.

Shanmugam, A. K., & Nesvizhskii, A. I. (2015). Effective leveraging of targeted search spaces for improving peptide identification in tandem mass spectrometry based proteomics. *Journal of Proteome Research, 14*(12), 5169–5178. doi:10.1021/acs.jproteome.5b00504.

Sheynkman, G. M., Shortreed, M. R., Frey, B. L., & Smith, L. M. (2013). Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Molecular & Cellular Proteomics, 12*(8), 2341–2353. doi:10.1074/mcp.O113.028142.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research, 19*(6), 1117–1123. doi:10.1101/gr.089532.108.

Sturm, M., & Kohlbacher, O. (2009). TOPPView: An open-source viewer for mass spectrometry data. *Journal of Proteome Research, 8*(7), 3760–3763. doi:10.1021/pr900171m.

Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research, 43*(12), e78. doi:10.1093/nar/gkv227.

Tay, A. P., Pang, C. N., Twine, N. A., Hart-Smith, G., Harkness, L., Kassem, M., & Wilkins, M. R. (2015). Proteomic validation of transcript isoforms, including those assembled from RNA-Seq data. *Journal of Proteome Research, 14*(9), 3541–3554. doi:10.1021/pr5011394.

Teleman, J., Rost, H. L., Rosenberger, G., Schmitt, U., Malmstrom, L., Malmstrom, J., & Levander, F. (2015). DIANA–algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics, 31*(4), 555–562. doi:10.1093/bioinformatics/btu686.

Ternent, T., Csordas, A., Qi, D., Gomez-Baena, G., Beynon, R. J., Jones, A. R., Hermjakob, H., & Vizcaino, J. A. (2014). How to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics, 14*(20), 2233–2241. doi:10.1002/pmic.201400120.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28*(5), 511–515. doi:10.1038/nbt.1621.

Trevisiol, S., Ayoub, D., Lesur, A., Ancheva, L., Gallien, S., & Domon, B. (2015). The use of proteases complementary to trypsin to probe isoforms and modifications. *Proteomics*. doi:10.1002/pmic.201500379.

Turewicz, M., & Deutsch, E. W. (2011). Spectra, chromatograms, metadata: mzML-the standard data format for mass spectrometer output. *Methods in Molecular Biology, 696*, 179–203. doi:10.1007/978-1-60761-987-1_11.

Tyanova, S., Temu, T., Carlson, A., Sinitcyn, P., Mann, M., & Cox, J. (2015). Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics, 15*(8), 1453–1456. doi:10.1002/pmic.201400449.

Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., & Martens, L. (2011). SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X! Tandem searches. *Proteomics, 11*(5), 996–999. doi:10.1002/pmic.201000595.

Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., Martens, L., & Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology, 33*(1), 22–24. doi:10.1038/nbt.3109.

Volders, P. J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J., & Mestdagh, P. (2013). LNCipedia: A database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research, 41*(Database issue), D246–D251. doi:10.1093/nar/gks915.

Volders, P. J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J., & Mestdagh, P. (2015). An update on LNCipedia: A database for annotated human lncRNA sequences. *Nucleic Acids Research, 43*(Database issue), D174–D180. doi:10.1093/nar/gku1060.

Walsh, C. T., Garneau-Tsodikova, S., & Gatto, G. J., Jr. (2005). Protein posttranslational modifications: The chemistry of proteome diversifications. *Angewandte Chemie International Edition, 44*(45), 7342–7372. doi:10.1002/anie.200501023.

Walzer, M., Qi, D., Mayer, G., Uszkoreit, J., Eisenacher, M., Sachsenberg, T., Gonzalez-Galarza, F. F., Fan, J., Bessant, C., Deutsch, E. W., Reisinger, F., Vizcaino, J. A., Medina-Aunon, J. A., Albar, J. P., Kohlbacher, O., & Jones, A. R. (2013). The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Molecular & Cellular Proteomics, 12*(8), 2332–2340. doi:10.1074/mcp.O113.028506.

Walzer, M., Pernas, L. E., Nasso, S., Bittremieux, W., Nahnsen, S., Kelchtermans, P., Pichler, P., van den Toorn, H. W., Staes, A., Vandenbussche, J., Mazanek, M., Taus, T., Scheltema, R. A., Kelstrup, C. D., Gatto, L., van Breukelen, B., Aiche, S., Valkenborg, D., Laukens, K., Lilley, K. S., Olsen, J. V., Heck, A. J., Mechtler, K., Aebersold, R., Gevaert, K., Vizcaino, J. A., Hermjakob, H., Kohlbacher, O., & Martens, L. (2014). qcML: An exchange format for quality control metrics from mass spectrometry experiments. *Molecular & Cellular Proteomics, 13*(8), 1905–1913. doi:10.1074/mcp.M113.035907.

Weisser, H., Nahnsen, S., Grossmann, J., Nilse, L., Quandt, A., Brauer, H., Sturm, M., Kenar, E., Kohlbacher, O., Aebersold, R., & Malmstrom, L. (2013). An automated pipeline for high-throughput label-free quantitative proteomics. *Journal of Proteome Research, 12*(4), 1628–1644. doi:10.1021/pr300992u.

Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., & Ma, B. (2012). PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics, 11*(4), M111 010587. doi:10.1074/mcp.M111.010587.

Identification of Small Novel Coding Sequences, a Proteogenomics Endeavor

4

Volodimir Olexiouk and Gerben Menschaert

Abstract

The identification of small proteins and peptides has consistently proven to be challenging. However, technological advances as well as multi-omics endeavors facilitate the identification of novel small coding sequences, leading to new insights. Specifically, the application of next generation sequencing technologies (NGS), providing accurate and sample specific transcriptome / translatome information, into the proteomics field led to more comprehensive results and new discoveries. This book chapter focuses on the inclusion of RNA-Seq and RIBO-Seq also known as ribosome profiling, an RNA-Seq based technique sequencing the +/− 30 bp long fragments captured by translating ribosomes. We emphasize the identification of micropeptides and neo-antigens, two distinct classes of small translation products, triggering our current understanding of biology. RNA-Seq is capable of capturing sample specific genomic variations, enabling focused neo-antigen identification. RIBO-Seq can identify translation events in small open reading frames which are considered to be non-coding, leading to the discovery of micropeptides. The identification of small translation products requires the integration of multi-omics data, stressing the importance of proteogenomics in this novel research area.

Volodimir Olexiouk and Gerben Menschaert equally contributed to the book chapter as first authors

V. Olexiouk (✉) • G. Menschaert
Lab of Bioinformatics and Computational Genomics
(BioBix), Faculty of Bioscience Engineering,
Department of Mathematical Modelling, Statistics
and Bioinformatics, Ghent University,
Coupure Links 653, Building A, Ghent 9000,
Belgium
e-mail: volodimir.olexiouk@ugent.be;
gerben.menschaert@ugent.be

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_4

Keywords
Mircopeptides • Small open reading frames • sORF • Neoantigens • Ribosome profiling • RIBO-Seq • Proteogenomics

4.1 Introduction

Unraveling protein biosynthesis is undoubtedly a multi-omics integration endeavor. From the DNA template (genomics) a region is transcribed (transcriptomics) and subsequently translated (translatomics) into protein products (proteomics). The aforementioned omics fields definitely intertwine, but are likewise considered self-sufficient, demonstrated by their vast complexity. Integration of matching multi-omics datasets, although challenging, can lead to more sound results and even new insights. Advances in bioinformatics have facilitated this multi-omics integration and expert tools became available to tackle specific parts of proteogenomics analyses (*e.g.*, PROTEOFORMER (Crappé et al. 2014a), PEPTIDESHAKER (Vaudel et al. 2015b), also see (Menschaert and Fenyö 2015) for a review of bioinformatics tools available in the proteogenomics field). An intriguing multi-omics empowered field tries to identify novel protein coding sequences. Direct assessment of proteins through mass spectrometry based proteomics analysis, combined with genomics, transcriptomics and translatomics information provides the necessary means to unravel the information flow from DNA to proteins (Wang and Zhang 2014). Particularly, the identification of micropeptides, translation products of small open reading frames, and neo-antigens, peptides resulting from proteins variants conceivably recognized by the immune system, are discussed in this book chapter. First, we will briefly describe the MS-based proteomics technology, highlighting the necessity for multi-omics integration in the research fields mentioned above.

As mentioned, the preferred methodology for protein / peptide identification is mass spectrometry (MS), a technique with high sensitivity and specificity (Cheng et al. 2014; Ryu 2014), capable of detecting up to 10,000 proteins from a single sample (Nagaraj et al. 2011). The global workflow in MS consists of enzymatic digestion of proteins extracted from the sample into peptides that are subsequently fragmented and analyzed by a mass spectrometer, providing peptide fragmentation spectra by registering the mass-to-charge ratio of ionized peptide fragments. Peptides are identified through database search engines (*e.g.*, X!tandem (Craig and Beavis 2004), Myrimatch (Tabb et al. 2007), MS-GF+ (Kim and Pevzner 2014; Granholm et al. 2014), Comet (Eng et al. 2015), MS Amanda (Dorfer et al. 2014)). A peptide-spectrum match (PSM) score is calculated by comparing experimental spectra against theoretical spectra, generated after *in silico* digestion of all proteins provided in a sequence database. Statistical validation methods in MS-based proteomics compute the false discovery rate (FDR) by means of a target-decoy approach assuming the reference database to contain the “true” pool of sequences represented in the sample (Hernandez et al. 2014). Consequently, deviation from this assumption impairs validation, implying that the main paradigm here is not to use the most exhaustive reference database, but to adversely focus on the most suitable reference database representing the true nature of the biological sample (Gupta et al. 2011; Nesvizhskii 2010; Wang et al. 2009a; Keller et al. 2002). Obviously, small proteins (micropeptides) produce less cleaved peptides and are often not present in reference protein databases, implicating their MS identification. Also, distinguishing resembling peptides can be complicated, as is frequently the case for neo-antigen identification.

Search engines and algorithms will definitely influence the peptide identification rate, but the reference database construction is pivotal, as inclusion is a prerequisite for identification.

Uniprot-KB (EMBL et al. 2013; Apweiler et al. 2014) is mostly used as the reference database in the MS-based proteomics identification process. This database is incomplete as it (partly) lacks information on novel proteoforms (isoforms), single nucleotide variation (SNV), indels (insertions and deletions), and gene fusion products. A more suitable reference database for novel protein identification is constructed containing all ORFs from the translation of the genome in its six reading frames. This strategy makes that all possible protein forms except for peptides spanning the exon junctions are included. That is why these are widely used for prokaryotes by virtue of a small genome and lack of splicing (Baudet et al. 2010). Since 98 % of the human genome is predicted to be non-coding (Lander et al. 2001), this approach would massively increase the search space resulting in an unattractive approach in terms of both computation time and error rate, while also omitting mutations, small open reading frames and non-AUG start sites.

Considering the 6 frame translation approach, only one sixth are true candidates, impairing the statistical validation model used (Hernandez et al. 2014; Blakeley et al. 2012). Furthermore, splice isoforms, single nucleotide variation and indels remain undetectable in a 6-frame translated reference database. A smaller reference database can be constructed from cDNA libraries or expressed sequence tags (EST), ensuring that the corresponding sequences are transcribed as they are derived from RNA (Hernandez et al. 2014). Furthermore, as the reference database has been constructed from RNA, alternative splice proteoforms may be included. Implementing such strategy in human has succeeded to compress the database to 3 % compared to a 6-frame reference database, with minimal sacrifices to the peptide sequence content (Edwards 2007). Another study using the Ensembl (Cunningham et al. 2014) database, including all isoforms, observed a 7 % increase in peptide identification compared to the non-redundant Swiss-Prot database (Fei et al. 2011). Tools as GENQUEST reduce the search space by filtering peptides on their mass and isoelectric point (Sevinsky et al. 2008). Although the afore-mentioned database choices have proven to be useful, the generated reference database contains sequences on a species wide level, where sample specific genomic (SNVs, indels) and RNA splice variations remain unregistered. Next generation sequencing (NGS) techniques enable the user to capture the transcriptome and/or translatome relatively accurate, fast and cost-efficient, thus enabling sample-specific reference database construction (Bahassi and Stambrook 2014). This review discusses how the integration of NGS techniques with MS-based proteomics enables the identification of novel, small proteins, strongly focusing on ribosome profiling and RNA-Seq. To illustrate the relevance of these techniques in current novel research fields, RNA-Seq mediated neo-antigen discovery and RIBO-Seq empowered micropeptide identification are discussed.

4.2 RNA-Seq

The majority of MS-based proteomic studies consist of comparing the obtained spectra against protein databases of known / predicted proteins, resulting in a high number of unidentified spectra. These unidentified spectra may map to novel peptides absent from the used protein database, represent splice variants, alternative open reading frames (*e.g.*, stop codon read-through, alternative start sites) or genetic variations (Ning and Nesvizhskii 2010). RNA-Seq provides a comprehensive profile of the transcriptome and enables the construction a database reflecting the native transcript composition, including those novel sequences (Woo et al. 2014; Marguerat and Bähler 2010; Wang et al. 2009b). A study performed by Wang et al. (2012) describes a workflow to derive a protein database from RNA-Seq data and records a substantial increase in peptide identifications in comparison to searches against an Ensembl database. Furthermore, RNA-Seq data allowed the detection of peptides containing SNPs associated with cancer. A workflow designed by Sheynkman et al. (2013), establishing a database focusing on splice junctions derived from RNA-Seq, identified unannotated transcript junctions from Jurkat cells. Compared to cDNA and EST libraries, RNA-Seq provides a more advanced and comprehensive methodology to identify novel splice junctions (Sheynkman et al. 2013). Moreover, RNA-Seq enables proteomics studies on non-model organisms with limited genome annotation (Lopez-Casado et al. 2012; Song et al. 2012; Armengaud 2013). Many RNA-Seq datasets are publically available (*e.g.*, in the Sequence Read Archive (Leinonen et al. 2011b) or European Nucleotide Archive (Leinonen et al. 2011a)) and can be utilized in proteogenomics applications. It is advised to pool multiple RNA-Seq experiments cumulatively (Woo et al. 2014) to construct a search space when non-matching proteomics and transcriptomics datasets are used.

4.2.1 Neo-antigens

The immune system recognizes an extensive range of antigens, which are distinguished as either 'self' or 'non-self' molecules. All human cells present peptide antigens on major histocompatibility complex (MHC) molecules, which interact with T-cell receptors (TCR), present on the plasma membrane of T-cells. When a peptide presented on the MHC is not recognized as 'self', this elicits a T-cell response, causing apoptosis or inactivation of the corresponding target cell. The presentation of 'non-self' peptide antigens may be induced by various reasons, ranging from viral infection to disturbed homeostasis (Singhal et al. 2013; Attaf et al. 2015). As tumor cells evolve from ordinary cells, they develop distinct characteristics recognizable by the immune system. Hence, the immune system is clearly of great importance in cancer development. The immune system can promote tumor growth by impairing tumor cell immunogenicity or act as a tumor suppressor by destroying or restraining tumor expansion (Koebel et al. 2007; Shankaran et al. 2001; Dunn et al. 2002). Immunotherapy, where T-cell activity is stimulated through the inhibition of the T-cell deactivation pathway (checkpoint blockade (Gubin et al. 2014)), has been shown to be an effective treatment in a variety of human malignancies (Wolchok and Chan 2014; Sharma and Allison 2015). For instance, Rosenberg (Hinrichs and Rosenberg 2014) demonstrated how infusion of tumor-infiltrating lymphocytes can be an effective treatment option in metastatic melanoma and antibody treatment sensitizing T-cell activation improved overall survival of metastatic melanoma patients (Hodi et al. 2010). The ability of T-cells to elicit a T-cell response based on the interaction with MHC molecules on tumor cells indicates the existence of tumor specific epitopes on antigens. These antigens can be derived from native proteins for which T-cell tolerance is incomplete (*e.g.,* tissue / time restricted proteins being expressed) or they can be formed from proteins absent from the human genome (*e.g.,* mutated proteins), called neo-antigens. Neo-epitopes are a product of tumor-specific DNA alterations and thus result in novel protein sequences (Schumacher and Schreiber 2015).

Studies in mouse models indicate that vaccination with neo-antigens increased tumor control in immunotherapy (Gubin et al. 2014; Yadav et al. 2014). However neo-antigen identification is tedious and limitations in MS sensitivity result in a substantial fraction of false negatives. Also, the identification of genomic variations in proteins does not guarantee MHC presentation. Combining transcriptomics sequencing techniques (RNA-Seq) to identify mutated proteins absent in native cells with proteomics identification of MHC presented antigens provides a feasible workflow useable in clinical studies. The global design of this workflow consists of the identification of tumor-specific genomic variation trough RNA-Seq, followed by an optional *in silico* filtering by algorithms to predict MHC antigen presentation and the construction of a database consisting of possible neo-antigen (Lu et al. 2014; Linnemann et al. 2014; Robbins et al. 2013). Next MS-based proteomics matches the experimentally identified MHC bound antigens against the RNA-Seq derived database, selecting high confidence neo-antigen. Functional essays can be performed to experimentally identify neo-antigens as demonstrated in mouse models, successfully treating cancer (Rizvi et al. 2015; Yadav et al. 2014; Bassani-Sternberg et al. 2015).

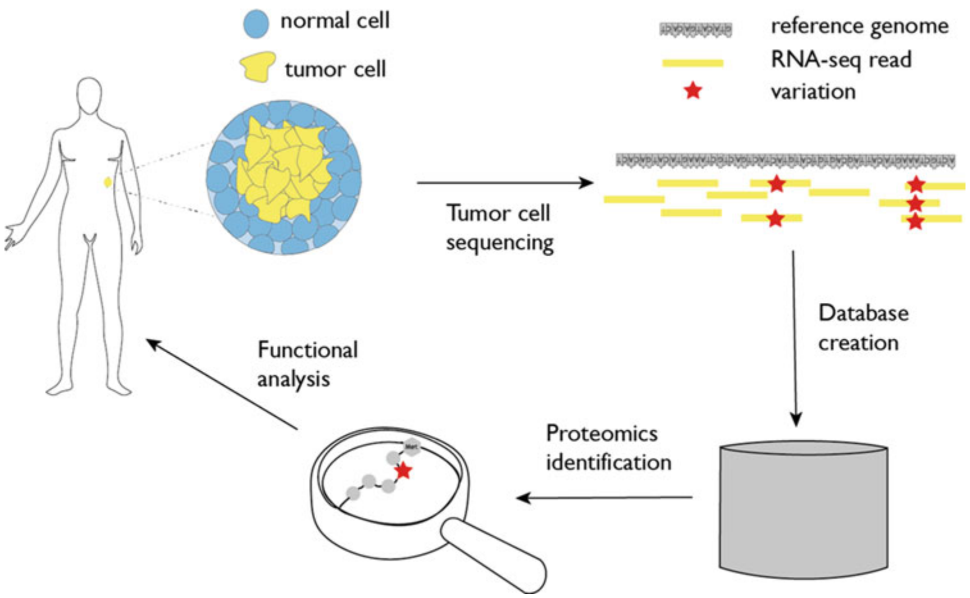


Fig. 4.1 A simplified neo-antigen identification workflow. Tumor cells are sequenced to identify genomic variations specific to these tumor cells, next a database is generated consisting of neo-antigen candidates. Optionally, *in silico* algorithms can be used to predict MHC antigen presentation, resulting in a more confident dataset. Next, MS-based proteomics identifies MHC bound antigens followed by functional analysis confirming candidate neo-antigens

Figure 4.1 provides a summary of the neo-antigen identification workflow.

4.3 RIBO-Seq

In the late 1960s, the ability of ribosomes to protect mRNA from endonuclease digestion was demonstrated (Steitz 1969). Despite this early discovery, it was not until the advent of NGS and the accompanying bioinformatics toolsets, that genome-wide translatome profiling became attainable. At the end of the twentieth century a technique named polysome profiling emerged (Johannes et al. 1999), yielding large scale analysis of translation. In summary, polysome profiling captures mRNA immobilized on translating ribosomes, separates these polyribosomes (*e.g.,* ultracentrifugation on a sucrose gradient) and subsequently sequences the obtained RNA fragments (Faye et al. 2014). This technique, identifying mRNA with ribosomal occupancy, saw various use-cases throughout the years and is still frequently applied (Piccirillo et al. 2014). However, it was with the advent of RIBO-Seq, enabling massive parallel sequencing of the +/− 30 nt mRNA fragments protected by ribosomes (RPFs), that in-depth assessment of the translatome was empowered (Ingolia et al. 2009, 2012, 2014). The main advantage of RIBO-Seq over polysome profiling is the ability to retrieve positional information obtained from these RPFs with sub-codon resolution, enabling accurate prediction of the ribosome A-site positions. The RIBO-Seq technique diverged into two complementary implementations, capturing either elongating ribosomes or initiating ribosomes. RIBO-Seq of elongating ribosomes is feasible through the addition of antibiotics inhibiting ribosome translocation (*e.g.,* cycloheximide (Ingolia et al. 2009) and emetine (Ingolia et al. 2012)), peptidyl transferase (*e.g.*, chloramphenicol) or by thermal freezing (Oh et al. 2011). Initiating ribosomes, allowing the deduction of translation initiation sites (TIS), is achieved through the addition of initiation blocking antibiotics (*e.g.,* harringtonine (Ingolia et al. 2012) or lactimidomycin (Lee et al. 2012)). Figure 4.2 sketches an overview of RIBO-Seq protocol.

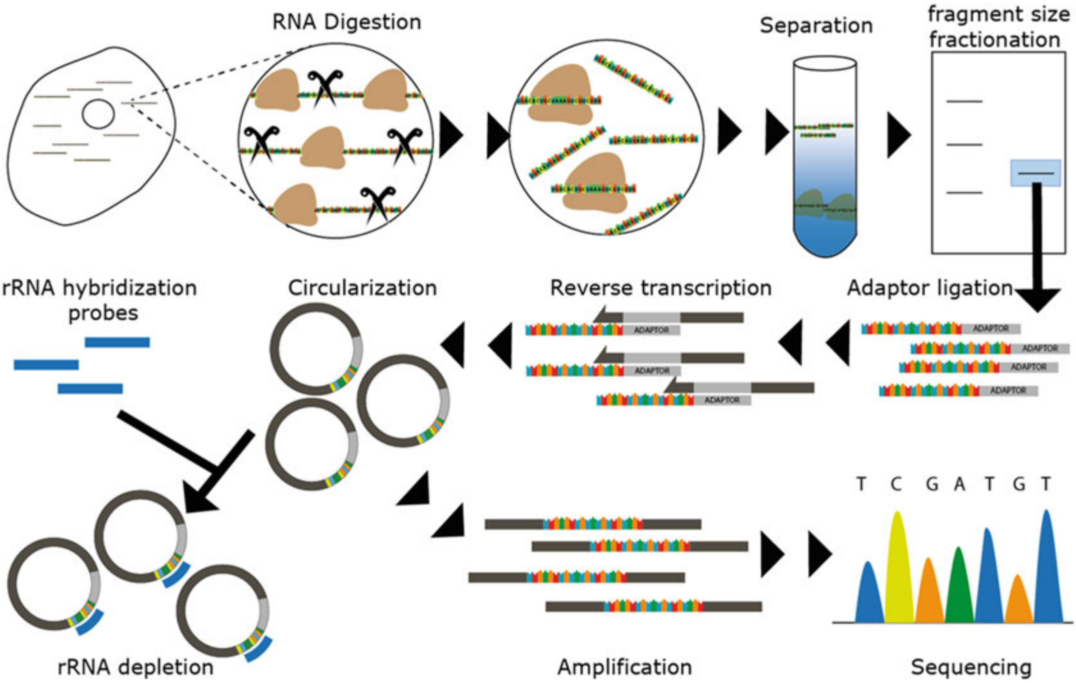


Fig. 4.2 A general overview of the RIBO-Seq protocol. First, cell lysates are prepared in conditions accurately reflecting in vivo translation. Secondly, addition of nucleases will digest RNA (nuclease footprinting), however the +/−30 nt mRNA fragments encapsulated by ribosomes are protected from digestion (ribosome footprints). Next, ribosome-footprints are separated from cell lysates followed by purification of ribosome protected RNA. Ligation of single-stranded adaptors enables reverse transcription. Subsequently, first strand reverse transcription products are circularized and transcript products hybridized to rRNA probes are depleted. Finally, PCR amplifies the remaining sequences that are subsequently sequenced. An in depth description of the protocol is provided by Ignolia et al. (2012b)

4.3.1 RIBO-Seq Unravels the Translatome

Although many variations are attributable to changes in gene transcripts, RIBO-Seq likewise reveals pervasive translational regulation (Michel and Baranov 2013). For example, Ignolia et al. (2009b) examined the ability of ribosome profiling to monitor changes in protein synthesis in response to starvation in yeast, observing translation changes in approximately one-third of the genes. Two other studies examining the translatome in response to heatshock (Shalgi et al. 2013) and proteotoxic stress (Liu et al. 2013) revealed interesting properties of the influence of chaperones on elongating ribosomes in response these stresses. In a study performed by Brar et al. (2012), exploring changes in expression during meiosis in yeast by performing RIBO-Seq over stage-specific time points, numerous dynamic events (including translation products of small open reading frames) were captured, unidentified by other techniques. A study performed by Stern-Ginossar et al. (2012) analyzed gene expression changes of human foreskin fibroblasts during cytomegalovirus infection. Measurements across different time-stamps revealed prominent viral gene translational regulation, where translation varied at least fivefold in 82 % of ORFs.

Furthermore, RIBO-Seq can identify novel translated regions, until now undetectable with other techniques. For instance several 5’-UTR ORFs, associated to a regulatory function (Ingolia et al. 2009, 2011; Brar et al. 2012), have been identified by ribosome profiling. The ORFs in 5’ untranslated regions are difficult to identify due to their specific characteristics: short length, limited coverage, non-AUG initiation, sometimes overlapping with canonical ORFs. Michel et al. (2012) demonstrated that given sufficient ribosome coverage, alternative reading frames are discernible by analyzing the triplet codon periodicity characteristic to translation and observable with the ribosome profiling technique. They reported on 5'-UTR ORFs with higher RPF intensity than the main canonical downstream ORF. In many cases these upstream ORFs (uORFs) partly overlapped with the canonical ORF. Furthermore Michel et al. identified frame transitions in translation, confirming well-known cases of frame shifts in humans. In a study performed by Gerashchenko et al. (2012) in yeast, four novel frame shift events were identified that correlated to oxidative stress. Also, the start site determination with the ribosome profiling technique enables the identification of ORFs with non-AUG start sites, resulting in numerous identified near-cognate initiation sites. Wan and Qian (2014) developed a database containing alterative translation initiation sites and their associated ORF identified by RIBO-Seq. Ribosomal activity was also observed in non-coding regions, revealing putative novel protein coding regions (Ingolia et al. 2012; Lee et al. 2012).

4.3.2 RIBO-Seq, a Bridge Between RNA-Seq and Proteomics

Protein inference from transcript abundance assumes constant RNA stability as well as stable translation rates. This assumption is erroneous as RNA stability can be highly variable and translation rates are volatile across transcripts. RIBO-Seq bridges the gap between RNA-Seq and proteomics by providing translational information, enabling improved inference from the transcriptome to the proteome and *vice versa*. RIBO-Seq is capable of detecting coding transcripts, but no direct evidence is provided whether these translated sequences ultimately yield stable protein products. Ribosomal occupancy could yield regulatory functions, but couls also point to unstable protein products or noise (Ingolia et al. 2014; Guttman and Rinn 2012). Several *in silico* tools and metrics were devised to predict the coding potential of ORFs (based on ribosome protected fragment length (Ingolia et al. 2014), triplet periodicity (Bazzini et al. 2014) and conservation (Lin et al. 2011)). However, MS-based validation remains a crucial confirmation technique in most cases. In turn, MS-based proteomics requires a database consisting of sample specific protein sequences. RIBO-Seq assisted database generation has several advantages over RNA-Seq generated databases. Novel proteoforms can be identified thus optimizing the search space (Calviello et al. 2015; Menschaert et al. 2013; Van Damme et al. 2014; Koch et al. 2014). This approach has been used by Fritsch et al. (2012) to identify 546 N-terminal protein extension in human, Menschaert et al. (2013) observed a 2.5 % increase in the overall protein identification rate using this approach. In a recent study performed by Fields et al. (2015), 1990 protein isoforms, 696 truncations, 341 extension and 1379 upstream ORFs were identified by RIBO-Seq. Automated pipelines facilitating RIBO-Seq integration in MS-based experiments, such as PROTEOFORMER (Crappé et al. 2014a), are readily available and easy to implement. Moreover Xie et al. (2015) developed an online database to query, analyze, visualize and download RIBO-Seq datasets.

4.4 Micropeptides

Micropeptides are defined as functional translation products originating from small open reading frames (sORFs). No consensus was reached regarding the sORF size and some studies consider an upper threshold of 200–250 codons (Hayden and Bosco 2008; Yang et al. 2011). However, the most widespread sORF size limit is 100 codons, a rule that we endorse here. A pioneering genome-wide study in 2003 on yeast suggested the functional importance of sORFs (Kessler et al. 2003), describing functionally conserved sORFs discovered by means of cross-species BLAST analysis. Only a few years later, Savard et al. (2006) identified mille-pattes in the red flour beetle by means of EST screening, a polycistronic peptide encoding four sORFs regulating HOX-genes. Kondo et al. (2007) and Galindo et al. (2007) examined mille-pattes analogs in *Drosophila melanogaster* resulting in the discovery of the tarsal-less (tal) and polished rice (pri) genes, respectively. This polycistronic mRNA, previously categorized as being non-coding, apparently was miss-annotated based on the ORFs size (Tupy et al. 2005). At the moment of writing, the *tal* and *pri* translation products are among the best characterized examples of micropeptides, regulating embryonic development throughout numerous insect species (Chanut-Delalande et al. 2014). The discovery of these *tal* and *pri* genes, together with the advent of ribosome profiling, boosted the research into sORF-encoded micropeptides. Several different research groups reported on the discovery of putatively coding sORFs using various techniques, pointing to novel functional micropeptides (Saghatelian and Couso 2015; Chu et al. 2015; Bazzini et al. 2014; Magny et al. 2013; Slavoff et al. 2013; Tonkin and Rosenthal 2015; Crappé et al. 2013; Pauli et al. 2014). Toddler, for example, is an embryonic signal that promotes cell movement (Pauli et al. 2014), Myoregulin regulates Ca^{2+} handling in muscle cells (Magny et al. 2013) and Sarcolipin regulates muscle-based thermogenesis in mammals (Tonkin and Rosenthal 2015). This is a relatively new research field (Crappé et al. 2014b; Andrews and Rothnagel 2014; Albuquerque et al. 2015), where the results of many *in silico* based studies and proteogenomics endeavors need further experimental validation.

4.4.1 In Silico Micropeptide Identification

Automated gene annotation systems correctly identify the majority of verified protein coding ORFs based on recognizable genomic sequence characteristics (*e.g.,* canonical initiation codons, splice sites, promoter sequences) (Sleator 2010). Most gene annotation algorithms set a lower threshold of 100 base triplets to exclude false positive annotations (Carninci et al. 2005; Frith et al. 2006a, b; Dinger et al. 2008). Recently, studies suggest that applying this lower threshold precludes the identification of numerous small proteins (Pauli et al. 2014; Bazzini et al. 2014; Ma et al. 2014; Frith et al. 2006a, b; Chng et al. 2013; Galindo et al. 2007; Crappé et al. 2013). Some computational approaches have been developed, such as uPEPperoni (Skarshewski et al. 2014) and sORFfinder (Hanada et al. 2009), providing *in silico* assessment of putatively coding sORFs, based on phylogenetic conservation. While the identification of sORFs is relatively straightforward, it does require a start and stop codon separated by at most 98 codons, the discrimination of coding vs. non-coding sORFs of this excessive pool of sORFs has proved to be more difficult. Due to their small size, many sORFs lacking any coding potential occur by chance. Cross-species conservation can be used as a proxy to function, but solely relying on phylogenetic conservation could prevent the identification of biologically relevant species-specific sORFs (Clamp et al. 2007). PhyloCSF (Lin et al. 2011) models phylogenetic relations between species by analyzing conservation at the amino acid level, rather than the nucleotide level and is most regularly used for small open reading frame assessment. It outperforms other methodologies (Reading Frame Conservation metrics, the regular CSF method or a d_n/d_s test) and is capable of identifying micropeptide coding sORFs as short as 13 amino acids (Guttman and Rinn 2012). Using mainly conservation as a criterion, Mackowiak et al. (2015) identified numerous conserved sORFs in different species (831 in *H. sapiens*, 350 in *M. musculus*, 211 in *D. rerio*, 194 in *D. melanogaster*, and 416 in *C. elegans*), some of which have been described and characterized previously.

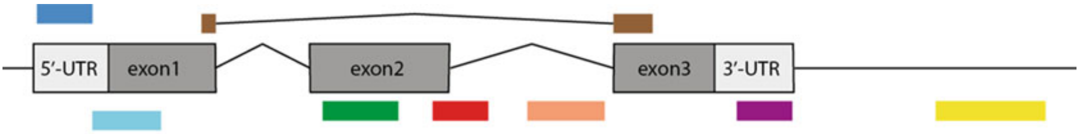
4.4.2 RIBO-Seq Enables the Identification of Translated sORFs

RNA-based transcriptomics is ignorant to ORF delineation; therefore most studies rely on conservation and pattern recognition for sORF identification. A recent study in yeast identified several micropeptides, one of which was also functionally characterized in influencing osmotic stress. The technique was based on using a 6-frame translation database derived from RNA-Seq data as a search space for subsequent MS fragmentation spectra matching (Yagoub et al. 2015). However, RNA-Seq does not indicate translation of the sORFs as opposed to RIBO-Seq. On top of pinpointing translated mRNA regions, RIBO-Seq can also reveal TIS, enabling the detection of non-AUG sORFs. *In silico* detection of non-AUG sORFs is laborious and difficult, since the search space becomes extensively larger, but from previous RIBO-Seq studies it has become clear that non-canonical start codons are more common than previously expected (Ingolia et al. 2011). Also, Slavoff et al. (2013) identified translation products from sORFs having non-AUG start sites using an MS-based proteogenomics approach. Recently, Fields et al. (2015) used a regression method on ribosome profiling data to identify sORFs that demonstrate an RPF length pattern and resemble that of annotated protein-coding ORFs. They discovered numerous sORFs, of which a subset shows very weak sequence conservation.

sORFs can be located in coding sequences (CDS), in 5'-untranslated regions (5'-UTR), in 3'-untranslated regions (3'-UTR), in intergenic regions (in-between genes) or in non-coding RNA regions. A first proof of 5'-UTR sORFs being translated was observed by Crowe et al. (2006). They revealed that 20 % of human 5'-UTR ORFs have TIS in an optimal Kozak sequence context, competent of ribosomal recognition. Follow-up studies revealed approximately 6750 conserved upstream TIS in mice (Lee et al. 2012) and approximately 3000 novel 5-UTR sORFs in human (Fritsch et al. 2012). A few 5'-UTR sORFs were identified encoding micropeptides (*e.g.*, MKKS in human (Akimoto et al. 2013), CPA1 in yeast (Werner et al. 1987)) with regulatory functions. Jorgenson (Jorgensen and Dorantes-Acosta 2012) claimed that 5'-UTR sORFs can regulate the downstream translation of the canonical ORF (also called the peptoswitch mechanism) as exemplified by CPA1. The discovery of dually coding transcripts (transcripts where more than one overlapping ORF can be translated), enabled the discovery of CDS-overlapping sORFs (*e.g.,* CASP1 (Ronsin et al. 1999) and altPrP (Vanderperre et al. 2011) in human). Most 3'-UTR sORFs are considered non-coding and are confirmed by the RIBO-Seq profiles that closely resemble those of non-coding ORFs. Still, a limited set of 3'-UTR sORFs was identified by MS-based techniques (*e.g.,* Bazzini et al. (2014) identified ten 3'-UTR sORFs using MS in combination with RIBO-Seq in a proteogenomics approach). Both sORFs in intergenic as well as in non-coding regions have been observed with RIBO-Seq (Lee et al. 2012). In particular, ribosomal activity on long non-coding RNA (lncRNA) fuelled a debate in the scientific community (Pauli et al. 2015) on whether or not lncRNAs are truly non-coding (Ruiz-Orera et al. 2014; Smith et al. 2014). Figure 4.3 provides an overview of sORFs identified in different (annotated) genomic regions.

4.4.3 Multi-omics Integration Is Still Indispensable

Ribosome occupancy does not necessarily mean translation into functional protein products; furthermore, RIBO-Seq is susceptible to noise. Besides conservation, several tools and metrics were developed to distinguish coding from non-coding sORFs. For example Ignolia et al. (2014) observed that the ribosome protected fragment (RPF) length distribution differs significantly between truly coding and non-coding ORFs and developed the FLOSS-score to distinguish between both categories (Fig. 4.4). Bazzini et al. (2014) developed the ORFscore, which calculates the preference of RPFs to accumulate in the first frame of coding sequences (Fig. 4.5), making full use of the triplet periodicity in the RIBO-Seq signal. The Ribosome Release Score (RRS) examines the release of translating ribosomes after hitting a stop codon (Guttman and Rinn 2012) (Fig. 4.4). More complex statistical methods are based on learning algorithms such as Coding Potential calculator (Kong et al. 2007), CRITICA (Badger and Olsen 1999), CSTMiner



Color	sORF location	Description
🟦	5′-UTR (no overlap)	5′-UTR sORF, located in the 5′-UTR region
🟦	5′-UTR (overlap)	5′-UTR sORF, located in the 5′-UTR region with exonic overlap
🟩	exonic	exonic sORF, located in an exonic region
🟫	exonic (alternatively spliced)	exonic sORF, located in an exonic region caused by alternative splicing
🟧	intronic (no overlap)	intronic sORF, located in an intronic region
🟥	intronic (overlap)	intronic sORF, located in an intronic region with exonic overlap
🟪	3′-UTR	3′-UTR sORF, located in the 3′-UTR region
🟨	intergenic	intergenic sORF, located in-between 2 genes

Fig. 4.3 sORFs classification. sORFs can be classified according to their genomic location, here an overview is provided of the different sORF classifications

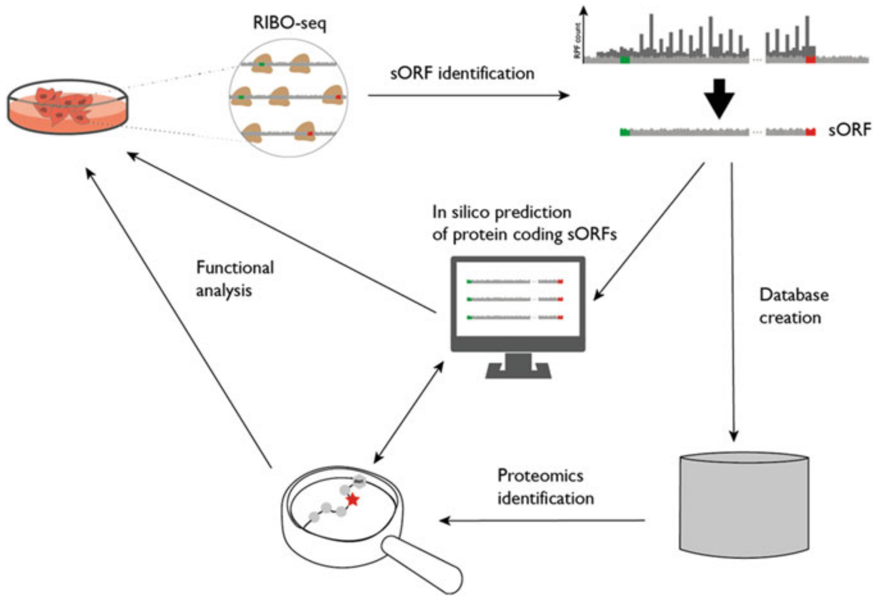


Fig. 4.4 Overview of coding potential assessment methods based on RIBO-seq. The FLOSS score compares the RPF-length distribution of sORFs with the RPF-length distribution of canonical protein-coding transcripts; strong disagreement between the two RPF-length distributions indicates non-coding behavior. The ORFscore calculates the preference of RPFs of coding ORFs to accumulate in the first frame of the coding sequence and the RRS provides a score based on the tendency of ribosome to dissociate from RNA after hitting a stop coding in coding ORFs

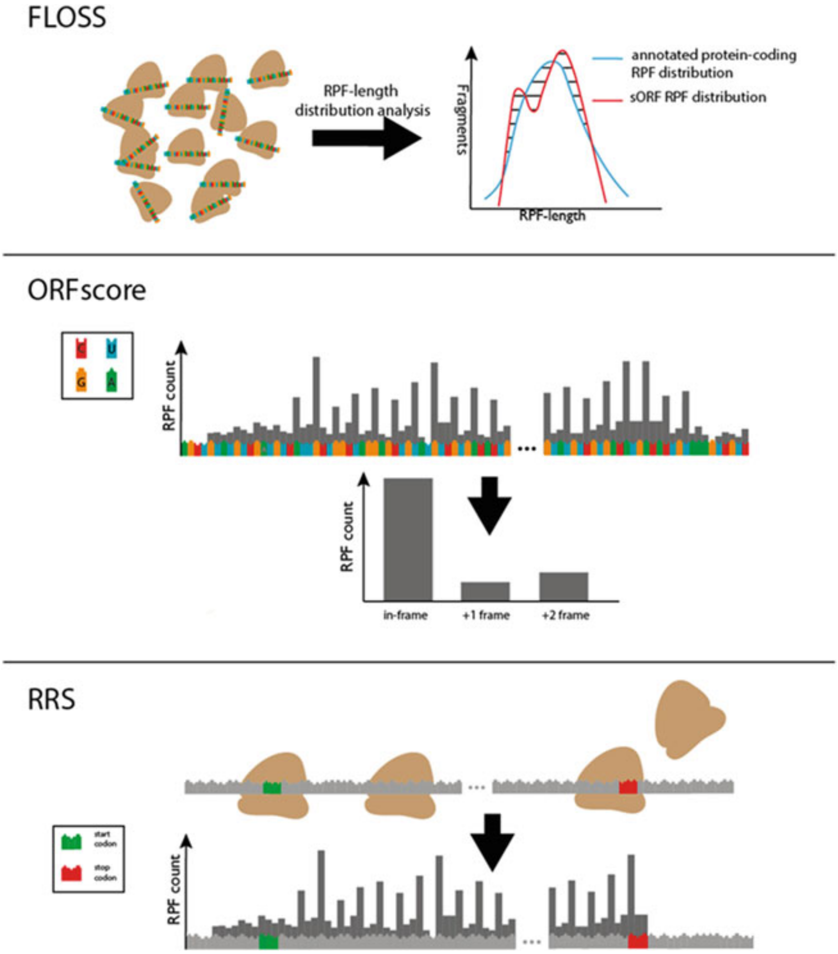


Fig. 4.5 A simplified micropeptide identification workflow. First, translating sORFs are identified using RIBO-seq. Next, candidate protein coding sORFs are predicting using methods described in the "Multi-omics integration is still indispensable" section and a database of translated sORFs is generated for proteomics identification. Results from both pathways can be combined in order to select micropeptides for functional analysis

(Castrignanò et al. 2004) and the recently described ORF-RATER (Fields et al. 2015) and RiboTaper (Calviello et al. 2015). ORF-RATER, a regression based translating ORF identifier based on RIBO-Seq data, discovered numerous novel ORFs, including sORFs with MS-evidence (Fields et al. 2015). Likewise, RiboTaper exploits a statistical approach to identify translated ORFs based on the nucleotide periodicity of RIBO-Seq data and correctly identified annotated protein coding sORFs, such as the aforementioned Toddler sORF (Calviello et al. 2015). However, in the novel field of micropeptide discovery, MS-based identification still remains indispensable. A proteogenomics approach generating a database of putatively coding sORFs derived from RIBO-Seq (or RNA-Seq) information, followed by MS-based proteomics identification creates an ideal setting for sORF discovery. Numerous sORFs have been identified using this approach (Ma et al. 2014; Bazzini et al. 2014; Mackowiak et al. 2015). A public database for sORFs (http://www.sorfs.org) exists, gathering multi-omics (RIBO-Seq and MS) evidence and *in silico* metrics. The resource currently harbors 266,342 sORFs across three model species (human, mouse, fruit fly) (Olexiouk et al. 2015), but will expand in the near future, with more data on other organism and cell types and including the latest "coding potential" metrics. Figure 4.4 provides an overview of the micropeptide identification workflow.

4.5 Conclusion and Future Perspectives

A multi-omics identification workflow for translation products is certainly advantageous, and is indispensable for novel (small) proteoform identifications. Such a proteogenomics approach is in many cases sample specific, enabling the analysis of sample specific variations. In cancer research, where variations obtained in a single cell may result in tumorous behavior and where these variations are frequently distinct between different tumor types, capturing such sample specific variations is crucial. Identification of neo-antigens in essence holds the identification of sample specific variation, obtainable by transcriptome sequencing technologies. However MS-based proteomics identification remains essential in order to perceive whether these transcript changes yield non-synonymous peptide variations. While still in its infancy, neo-antigen research increases the overall understanding of the immune system and moreover holds important therapeutic value.

The RIBO-Seq enabled genome-wide assessment of translation (translatomics) bridges two omics fields: transcriptomics and proteomics. Genome wide analysis of this ribosome profiling information already resulted in the identification of numerous sORFs with coding potential, questioning the non-coding character of sORFs. Follow-up analyses observed sORFs that resemble canonical coding ORFs and some are in the mean fully characterized as being coding. Over the last years, various tools and metrics were devised to assess the coding potential of sORFs (both conservation and sequence based). Also, workflows aiding the integration of RIBO-Seq information and MS-based proteomics are becoming available, *e.g.*, PROTEOFORMER (Crappé et al. 2014a). The scientific community is becoming aware of sORFs as potentially protein coding units. As a result, public sORF databases, such as http://www.sorfs.org, will be highly useful in the experimental design of future experiments (Olexiouk et al. 2015). Moreover, already conducted experiments (with an emphasis on MS-based proteomics studies) must be reprocessed to account for micropeptides. The scientific community is becoming aware of the large amount of publically available proteomics data accumulated over the past years that is currently being left untouched, while our scientific knowledge and technology evolved tremendously (Vaudel et al. 2015a, b; Verheggen et al. 2015). The sORFs.org database already holds a pilot study where 1172 publically available MS datasets from PRIDE were reprocessed, providing MS-evidence for more than 5000 micropeptides. Cumulative evidence that sORFs are able to encode functional micropeptides has been gathered, but their exact biological relevance often remains to be determined. Undoubtedly, future research on overexpression or knock-down will reveal more about the functional roles of specific sORF-encoded micropeptides.

4.6 Funding

Postdoctoral Fellows of the Research Foundation – Flanders (FWO-Vlaanderen) [G.M.,12A7813N]. Research Foundation – Flanders (FWO-Vlaanderen) [V.O, G0D3114N].

Conflict of Interest Statement None declared.

References

Akimoto, C., et al. (2013). Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochimica et Biophysica Acta - General Subjects, 1830*(3), 2728–2738.

Albuquerque, J. P., Tobias-santos, V., & Rodrigues, A. C. (2015). small ORFs: A new class of essential genes for development. *Genetics and Molecular Biology, 283*, 278–283.

Andrews, S. J., & Rothnagel, J. a. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics, 15*(3), 193–204.

Apweiler, R., et al. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research, 42*(D1), D191–D198.

Armengaud, J. (2013). Microbiology and proteomics, getting the best of both worlds! *Environmental Microbiology, 15*(1), 12–23.

Attaf, M., et al. (2015). The T cell antigen receptor: The Swiss Army knife of the immune system. *Clinical & Experimental Immunology, 181*(1), 1–18.

Badger, J. H., & Olsen, G. J. (1999). CRITICA: Coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution, 16*(4), 512–524.

Bahassi, E. M., & Stambrook, P. J. (2014). Next-generation sequencing technologies: Breaking the sound barrier of human genetics. *Mutagenesis, 29*(5), 303–310.

Bassani-Sternberg, M., et al. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Molecular & Cellular Proteomics, 14*(3), 658–673.

Baudet, M., et al. (2010). Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Molecular & Cellular Proteomics, 9*(2), 415–426.

Bazzini, A. A., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO Journal, 33*(9), 981–993.

Blakeley, P., Overton, I. M., & Hubbard, S. J. (2012). Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *Journal of Proteome Research, 11*(11), 5221–5234.

Brar, G. a., et al. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science, 335*(6068), 552–557.

Calviello, L. et al. (2015, December). Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, *13*(2), 1–9.

Carninci, P., et al. (2005). The transcriptional landscape of the mammalian genome. *Science, 309*(5740), 1559–1563.

Castrignanò, T. et al. (2004). CSTminer: A web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. *Nucleic Acids Research*, *32*(Web Server issue), W624–W627.

Chanut-Delalande, H., et al. (2014). Pri peptides are mediators of ecdysone for the temporal control of development. *Nature Cell Biology, 16*(11), 1035–1044.

Cheng, K., et al. (2014). Fit-for-purpose curated database application in mass spectrometry-based targeted protein identification and validation. *BMC Research Notes, 7*, 444.

Chng, S. C., et al. (2013). ELABELA: A hormone essential for heart development signals via the apelin receptor. *Developmental Cell, 27*(6), 672–680.

Chu, Q., Ma, J., & Saghatelian, A. (2015). Identification and characterization of sORF-encoded polypeptides. *Critical Reviews in Biochemistry and Molecular Biology, 50*(2), 134–141.

Clamp, M., et al. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America, 104*(49), 19428–19433.

Craig, R., & Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics, 20*(9), 1466–1467.

Crappé, J., et al. (2013). Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics, 14*, 648.

Crappé, J., Ndah, E., et al. (2014a). PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research, 10*, 1–10.

Crappé, J., Van Criekinge, W., & Menschaert, G. (2014b). Little things make big things happen: A summary of micropeptide encoding genes. *EuPA Open Proteomics, 3*, 128–137.

Crowe, M. L., Wang, X.-Q., & Rothnagel, J. a. (2006). Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics, 7*, 16.

Cunningham, F., et al. (2014). Ensembl 2015. *Nucleic Acids Research, 43*(D1), D662–D669.

Dinger, M. E., et al. (2008). Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Computational Biology, 4*(11), e1000176.

Dorfer, V., et al. (2014). MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research, 13*(8), 3679–3684.

Dunn, G. P., et al. (2002). Cancer immunoediting: From immunosurveillance to tumor escape. *Nature Immunology, 3*(11), 991–998.

Edwards, N. J. (2007). Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Molecular Systems Biology, 3*(1), 102.

EMBL, SIB Swiss Institute of Bioinformatics, & Protein Information Resource (PIR). (2013). UniProt. *Nucleic Acids Research, 41*, D43–D47.

Eng, J. K., et al. (2015). A deeper look into comet—Implementation and features. *Journal of The American Society for Mass Spectrometry, 26*(11), 1865–1874.

Faye, M. D., Graber, T. E., & Holcik, M. (2014). Assessment of selective mRNA translation in mammalian cells by polysome profiling. *Journal of Visualized Experiments, 92*, 1–8.

Fei, S. S., et al. (2011). Protein database and quantitative analysis considerations when integrating genetics and proteomics to compare mouse strains. *Journal of Proteome Research, 10*(7), 2905–2912.

Fields, A. P., et al. (2015). A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Molecular Cell, 60*(5), 816–827.

Frith, M. C., et al. (2006a). Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biology, 3*(1), 40–48.

Frith, M. C., et al. (2006b). The abundance of short proteins in the mammalian proteome. *PLoS Genetics, 2*(4), 515–528.

Fritsch, C., et al. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Research, 22*(11), 2208–2218.

Galindo, M. I., et al. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biology, 5*(5), 1052–1062.

Gerashchenko, M. V., Lobanov, a. V., & Gladyshev, V. N. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proceedings of the National Academy of Sciences, 109*(43), 17394–17399.

Granholm, V., et al. (2014). Fast and accurate database searches with MS-GF+Percolator. *Journal of Proteome Research, 13*(2), 890–897.

Gubin, M. M., et al. (2014). Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature, 515*(7528), 577–581.

Gupta, N., et al. (2011). Target-decoy approach and false discovery rate: When things may go wrong. *Journal of The American Society for Mass Spectrometry, 22*(7), 1111–1120.

Guttman, M., & Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature, 482*(7385), 339–346.

Hanada, K., et al. (2009). sORF finder: A program package to identify small open reading frames with high coding potential. *Bioinformatics, 26*(3), 399–400.

Hayden, C. a., & Bosco, G. (2008). Comparative genomic analysis of novel conserved peptide upstream open reading frames in Drosophila melanogaster and other dipteran species. *BMC Genomics, 9*, 61.

Hernandez, C., Waridel, P., & Quadroni, M. (2014). Database construction and peptide identification strategies for proteogenomic studies on sequenced genomes. *Current Topics in Medicinal Chemistry, 14*(3), 425–434.

Hinrichs, C. S., & Rosenberg, S. a. (2014). Exploiting the curative potential of adoptive T-cell therapy for cancer. *Immunological Reviews, 257*(1), 56–71.

Hodi, F. S., et al. (2010). Improved survival with ipilimumab in patients with metastatic melanoma. *The New England Journal of Medicine, 363*(8), 711–723.

Ingolia, N. T. et al. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, *324*(5924), 218–223.

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell, 147*(4), 789–802.

Ingolia, N. T., et al. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols, 7*(8), 1534–1550.

Ingolia, N. T., et al. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports, 8*(5), 1365–1379.

Johannes, G., et al. (1999). Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proceedings of the National Academy of Sciences of the United States of America, 96*(23), 13118–13123.

Jorgensen, R. A., & Dorantes-Acosta, A. E. (2012, August). Conserved peptide upstream open reading frames are associated with regulatory genes in Angiosperms. *Frontiers in Plant Science*, *3*, 1–11.

Keller, A., et al. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry, 74*(20), 5383–5392.

Kessler, M. M., et al. (2003). Systematic discovery of new genes in the Saccharomyces cerevisiae genome. *Genome Research, 13*(2), 264–271.

Kim, S., & Pevzner, P. a. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications, 5*, 5277.

Koch, A., et al. (2014). A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics, 14*, 2688–2698.

Koebel, C. M., et al. (2007). Adaptive immunity maintains occult cancer in an equilibrium state. *Nature, 450*(7171), 903–907.

Kondo, T., et al. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology, 9*(6), 660–665.

Kong, L. et al. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, *35*(Web Server issue), W345–W349.

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

Lee, S. S., et al. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America, 109*(37), E2424–E2432.

Leinonen, R., Akhtar, R., et al. (2011a). The European nucleotide archive. *Nucleic Acids Research, 39*(Database issue), D28–D31.

Leinonen, R., Sugawara, H., & Shumway, M. (2011b). The sequence read archive. *Nucleic Acids Research, 39*(Database issue), D19–D21.

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics, 27*(13), 275–282.

Linnemann, C., et al. (2014). High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nature Medicine, 21*(1), 81–85.

Liu, B., Han, Y., & Qian, S. B. (2013). Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. *Molecular Cell, 49*(3), 453–463.

Lopez-Casado, G., et al. (2012). Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case. *Proteomics, 12*, 761–774.

Lu, Y. C., et al. (2014). Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions. *Clinical Cancer Research, 20*(13), 3401–3410.

Ma, J., et al. (2014). Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *Journal of Proteome Research, 13*(3), 1757–1765.

Mackowiak, S. D., et al. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology, 16*(1), 179.

Magny, E. G. et al. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, *341*(6150), 1116–1120.

Marguerat, S., & Bähler, J. (2010). RNA-Seq: From technology to biology. *Cellular and Molecular Life Sciences, 67*(4), 569–579.

Menschaert, G., & Fenyö, D. (2015). Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrometry Reviews, 34*(1), 16.

Menschaert, G., et al. (2013). Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & Cellular Proteomics, 12*(7), 1780–1790.

Michel, A. M., & Baranov, P. V. (2013). Ribosome profiling: A Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdisciplinary Reviews: RNA, 4*(5), 473–490.

Michel, A. M., et al. (2012). Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Research, 22*(11), 2219–2229.

Nagaraj, N., et al. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology, 7*(548), 1–8.

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics, 73*(11), 2092–2123.

Ning, K., & Nesvizhskii, A. I. (2010). The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: A preliminary assessment. *BMC Bioinformatics*, *11*(Suppl 11), S14.

Oh, E., et al. (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell, 147*(6), 1295–1308.

Olexiouk, V. et al. (2015). sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research*, p.gkv1175.

Pauli, A. et al. (2014). Toddler: An embryonic signal that promotes cell movement via Apelin receptors. *Science (New York, N.Y.)*, *343*(6172), 1248636.

Pauli, A., Valen, E., & Schier, A. F. (2015). Identifying (non-)coding RNAs and small peptides: Challenges and opportunities. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology, 37*(1), 103–112.

Piccirillo, C. a., et al. (2014). Translational control of immune responses: From transcripts to translatomes. *Nature Immunology, 15*(6), 503–511.

Rizvi, N. A., et al. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science, 348*(6230), 124–128.

Robbins, P. F., et al. (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nature Medicine, 19*(6), 747–752.

Ronsin, C. et al. (1999). A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ. *Journal of Immunology (Baltimore, Md. : 1950)*, *163*(1), 483–490.

Ruiz-Orera, J., et al. (2014). Long non-coding RNAs as a source of new peptides. *eLife, 3*, e03523.

Ryu, S. Y. (2014). Bioinformatics tools to identify and quantify proteins using mass spectrometry data. *Advances in Protein Chemistry and Structural Biology, 94*, 1–17.

Saghatelian, A., & Couso, J. P. (2015). Discovery and characterization of smORF-encoded bioactive polypeptides. *Nature Chemical Biology, 11*(12), 909–916.

Savard, J., et al. (2006). A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell, 126*(3), 559–569.

Schumacher, T. N., & Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science (New York, N.Y.)*, *348*(6230), 69–74.

Sevinsky, J. R., et al. (2008). Whole genome searching with shotgun proteomic data: Applications for genome annotation. *Journal of Proteome Research, 7*(1), 80–88.

Shalgi, R., et al. (2013). Widespread regulation of translation by elongation pausing in heat shock. *Molecular Cell, 49*(3), 439–452.

Shankaran, V., et al. (2001). IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature, 410*(6832), 1107–1111.

Sharma, P., & Allison, J. P. (2015). The future of immune checkpoint therapy. *Science (New York, N.Y.)*, *348*(6230), 56–61.

Sheynkman, G. M., et al. (2013). Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Molecular & Cellular Proteomics, 12*(8), 2341–2353.

Singhal, A., Mori, L., & De Libero, G. (2013). T cell recognition of non-peptidic antigens in infectious diseases. *The Indian Journal of Medical Research, 138*(5), 620–631.

Skarshewski, A., et al. (2014). uPEPperoni: An online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics, 15*, 36.

Slavoff, S. a., et al. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology, 9*(1), 59–64.

Sleator, R. D. (2010). An overview of the current status of eukaryote gene prediction strategies. *Gene, 461*(1–2), 1–4.

Smith, J. E., et al. (2014). Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Reports, 7*(6), 1858–1866.

Song, J., et al. (2012). An improvement of shotgun proteomics analysis by adding next-generation sequencing transcriptome data in orange. *PloS One, 7*(6), 5–10.

Steitz, J. a. (1969). Nucleotide sequences of the ribosomal binding sites of bacteriophage R17 RNA. *Cold Spring Harbor Symposia on Quantitative Biology, 34*, 621–630.

Stern-Ginossar, N. et al. (2012). Decoding human cytomegalovirus. *Science (New York, N.Y.), 338*(6110), 1088–1093.

Tabb, D. L., Fernando, C. G., & Chambers, M. C. (2007). MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research, 6*(2), 654–661.

Tonkin, J., & Rosenthal, N. (2015). One small step for muscle: A new micropeptide regulates performance. *Cell Metabolism, 21*(4), 515–516.

Tupy, J. L., et al. (2005). Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America, 102*(15), 5495–5500.

Van Damme, P., et al. (2014). N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Molecular & Cellular Proteomics, 13*(5), 1245–1261.

Vanderperre, B., et al. (2011). An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *The FASEB Journal, 25*(7), 2373–2386.

Vaudel, M., & Verheggen, K. et al. (2015). Exploring the potential of public proteomics data. *Proteomics*, (January 2016), 1–30.

Vaudel, M., Burkhart, J. M., et al. (2015b). PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology, 33*(1), 22–24.

Verheggen, K. et al. (2015). Pladipus enables universal distributed computing in proteomics bioinformatics. *Journal of Proteome Research*, p.acs. jproteome.5b00850.

Wan, J., & Qian, S. B. (2014). TISdb: A database for alternative translation initiation in mammalian cells. *Nucleic Acids Research*, *42*(November 2013), 845–850.

Wang, X., & Zhang, B. (2014). Integrating genomic, transcriptomic, and interactome data to improve peptide and protein identification in shotgun proteomics. *Journal of Proteome Research, 13*(6), 2715–2723.

Wang, G., et al. (2009a). Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Analytical Chemistry, 81*(1), 146–159.

Wang, Z., Gerstein, M., & Snyder, M. (2009b). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57–63.

Wang, X., et al. (2012). Protein identification using customized protein sequence databases derived from RNA-Seq data. *Journal of Proteome Research, 11*(2), 1009–1017.

Werner, M., et al. (1987). The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression. *Cell, 49*(6), 805–813.

Wolchok, J., & Chan, T. (2014). Cancer: Antitumour immunity gets a boost. *Nature, 515*, 496–498.

Woo, S., et al. (2014). Proteogenomic database construction driven from large scale RNA-Seq data. *Journal of Proteome Research, 13*(1), 21–28.

Xie, S.-Q. et al. (2015). RPFdb: A database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Research*, p.gkv972.

Yadav, M., et al. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature, 515*(7528), 572–576.

Yagoub, D. et al. (2015). Proteogenomic discovery of a small, novel protein in yeast reveals a strategy for the detection of unannotated short open reading frames. *Journal of Proteome Research*, p.acs.jproteome.5b00734.

Yang, X., et al. (2011). Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Research, 21*(4), 634–641.

Using Proteomics Bioinformatics Tools and Resources in Proteogenomic Studies

5

Marc Vaudel, Harald Barsnes, Helge Ræder, and Frode S. Berven

Abstract

Proteogenomic studies ally the omic fields related to gene expression into a combined approach to improve the characterization of biological samples. Part of this consists in mining proteomics datasets for non-canonical sequences of amino acids. These include intergenic peptides, products of mutations, or of RNA editing events hypothesized from genomic, epigenomic, or transcriptomic data. This approach poses new challenges for standard peptide identification workflows. In this chapter, we present the principles behind the use of peptide identification algorithms and highlight the major pitfalls of their application to proteogenomic studies.

Keywords

Proteogenomics • Proteomics • Bioinformatics

M. Vaudel (✉)
Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway

Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway
e-mail: marc.vaudel@uib.no

H. Barsnes
Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway

H. Ræder
KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway

Department of Pediatrics, Haukeland University Hospital, Bergen, Norway

F.S. Berven
Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_5

5.1 Using Proteomics Bioinformatics Tools and Resources in Proteogenomic Studies

Proteomics aims at characterizing the entire set of proteins present in a sample at a given time point. The most encountered approach is to digest proteins into peptides prior to analysis by liquid chromatography (LC) coupled to tandem mass spectrometry (MS) (Aebersold and Mann 2003). Specialized bioinformatic tools are in turn used to infer peptide sequences and eventually post-translational modifications (PTMs), and from there infer a list of proteins present in the sample with their respective abundances (Altelaar et al. 2013).

Three main approaches are available for the identification of peptides from tandem mass spectra, as reviewed in (Vaudel et al. 2012b): (i) spectral library searching, (ii) *de novo* sequencing, and (iii) sequence databases searching. In the first, experimental spectra are compared to a library of already identified spectra, and the quality of the matches is evaluated to infer the presence of known peptides. In the second, a sequence or partial sequence is inferred from the spectrum by comparing the distance between peaks to possible amino acids and PTM masses. In the third, theoretical spectra are derived from a database of expected protein sequences, and the match between experimental and theoretical spectra is in turn evaluated to infer the presence of peptides.

Importantly, in most cases, a reference protein sequence database is used to map the peptide sequences back to the protein level, a task called protein inference (Nesvizhskii and Aebersold 2005). Given that spectral library searching can only find peptides which have been previously measured, it is generally not used for discovery studies but mainly for targeted protein quantification (Domon and Aebersold 2006), and despite constant progresses in instrumentation, the information contained in spectra is generally insufficient for *de novo* sequencing to compete with sequence database searching. As a result, sequence database searching tools, so-called search engines, are by far the most encountered approach for peptide identification.

5.2 Database Search Engines

Proteomic search engines take as input (i) peak lists from mass spectrometry experiments, (ii) expected protein sequences, and (iii) user defined search parameters. The processing of raw mass spectrometry data prior to identification typically includes signal processing to denoise spectra, reduce the baseline, and perform peak picking, *i.e.*, transform linear spectra of bell-shaped peaks, called profile spectra, into discrete peak lists, called centroided spectra (Lange et al. 2006; Vaudel et al. 2010). However, with the advent of high resolution mass spectrometers including built-in signal processing units, this task has been dramatically simplified, and the raw data can often simply be used without additional processing.

The conversion of raw mass spectrometry files to open formats can be easily conducted by msconvert as part of the ProteoWizard package (Kessner et al. 2008), and if needed, signal processing methods such as peak picking can be applied by ProteoWizard (French et al. 2015) or within the OpenMS package (Kohlbacher et al. 2007).

Various public resources with canonical protein sequences are available, some specialized in terms of organisms, diseases or sub proteomes, while others, generalist, attempt at providing comprehensive collections of all known sequences. The main generalist protein databases are the UniProt knowledgebase (Apweiler et al. 2004), the closely connected Ensembl database (Hubbard et al. 2002; Yates et al. 2016), the National Center for Biotechnology Information (NCBI) Protein database (Coordinators 2016) (accessible *via* the Entrez Global Query system providing sequences from multiple sources including the Reference Sequence (RefSeq) Database (Pruitt et al. 2005)), and the DNA Data Bank of Japan (DDBJ) (Tateno et al. 2002).

Through the search parameters, the user designs the search space of the algorithm, *i.e.*, peptide and fragment ions that are expected from the given theoretical protein sequences, and sets the degrees of freedom with which the algorithm matches experimental mass spectra to these theoretical ions. The parameters typically include: (i) information on how proteins are expected to be cleaved with a maximum limit on the number of missed cleavages, (ii) the expected fragmentation behaviour of peptides and resulting ions, (iii) mass tolerances, and (iv) expected PTMs. Depending on their prevalence, PTMs can be systematically accounted for, so-called *fixed* or *static* modifications, or the algorithm can iterate the possible modification statuses of a peptide, so-called *variable* or *dynamic* modifications. A detailed description on how to tune the role of search engines can be found in the CompOmics Tutorials (Vaudel et al. 2014b) (compomics.com/bioinformatics-for-proteomics).

Taking these settings into account, the algorithm will compare the expected fragmentation pattern of each possible peptide to each experimental spectrum, and, for every spectrum, return a list of so-called peptide spectrum matches (PSMs) along with a score of each match. The larger the search space, the more complex this task becomes, and, as detailed below, managing the search space is one of the main challenges when using search engines in proteogenomic studies.

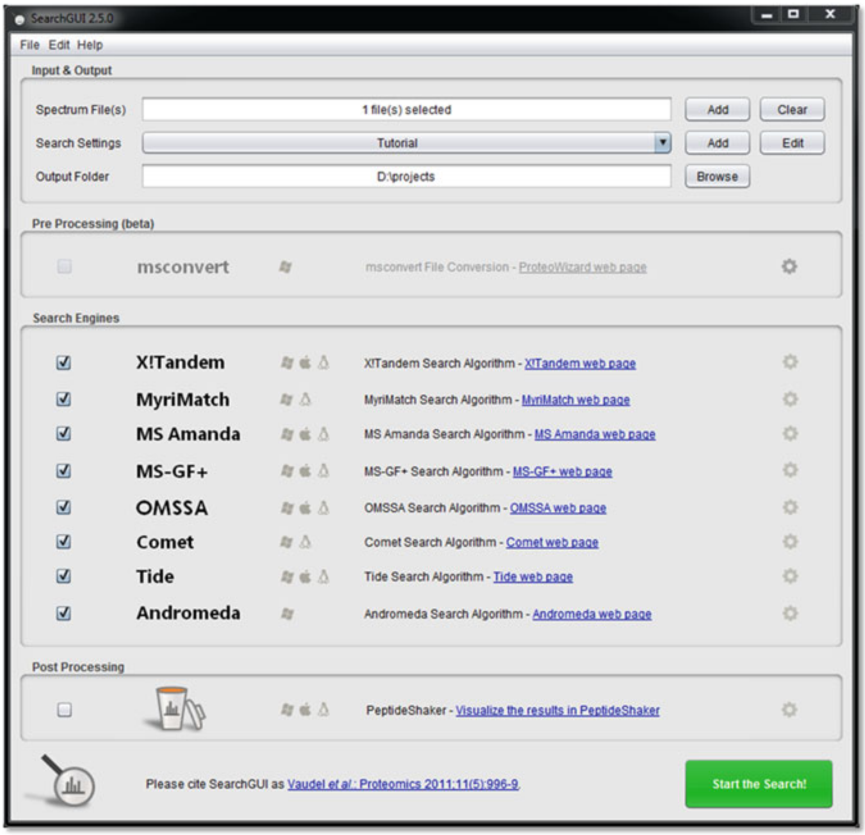
Numerous algorithms were developed over the past decades, most of them listed at the 'OMIC tools' web site (Henry et al. 2014) (omictools.com/database-search-category). Some of them have been included into generic proteomic software suites like the Trans Proteomics Pipeline (TPP) (Deutsch et al. 2010), OpenMS (Kohlbacher et al. 2007), and MaxQuant (Cox and Mann 2008). Alternatively, a simple way to operate search engines is to use them *via* SearchGUI (Vaudel et al. 2011a), a user-friendly graphical and command line interface making it possible to harness multiple search engines – at time of writing, X!Tandem (Craig and Beavis 2004), MyriMatch (Tabb et al. 2007), MS Amanda (Dorfer et al. 2014), MS-GF+ (Kim and Pevzner 2014), OMSSA (Craig et al. 2004), Comet (Eng et al. 2013), Tide (Diament and Noble 2011), and Andromeda (Cox et al. 2011) – see Fig. 5.1.

5.3 Results Integration

Upon completion of the search, result files are produced by each search engine, containing lists of PSMs. These are then parsed and assembled into proteins using post processing software tools. Some of the critical tasks of such tools are to select a representative hit per spectrum, infer proteins from peptides (Nesvizhskii and Aebersold 2005), and estimate error rates. As discussed in detail previously (Nesvizhskii 2010; Ma et al. 2012), two methods exist for the evaluation of error rates, the modelling of score distributions (Keller et al. 2002) or the so-called target-decoy strategy (Elias and Gygi 2010). In the latter, artificial decoy sequences are added to the sequence database or searched in parallel. The distribution of the decoy scores is then used to model the distribution of the scores of false positive hits to derive error rates. These can be estimated locally, providing a probability for an identified compound to be a false result, a posterior error probability (PEP), or for the entire dataset, providing a global false discovery rate (FDR). In addition to the statistical evaluation of error rates, it is possible to do manual or semi-manual inspection of the results (Helsens et al. 2008).

When operating outside the software frameworks mentioned above, a simple way to aggregate the results of multiple search engines is to process them with PeptideShaker (Vaudel et al. 2015), a user friendly tool designed to work seamlessly in combination with SearchGUI. Additionally to peptide to protein mapping and protein inference, error rates and matches quality control, PeptideShaker performs, quality filtering, and PTM localization, all *via* an intuitive and interactive graphical user interface, see Fig. 5.2.

Upon completion of the processing, the interpretation of proteomics results strongly relies on their biological and functional contextualization (Vaudel et al. 2014a). The data can, for example,



Input & Output

Spectrum File(s): 1 file(s) selected

Search Settings: Tutorial

Output Folder: D:\projects

Pre Processing (beta)

□ msconvert — msconvert File Conversion - ProteoWizard web page

Search Engines

☑ X!Tandem — X!Tandem Search Algorithm - X!Tandem web page

☑ MyriMatch — MyriMatch Search Algorithm - MyriMatch web page

☑ MS Amanda — MS Amanda Search Algorithm - MS Amanda web page

☑ MS-GF+ — MS-GF+ Search Algorithm - MS-GF+ web page

☑ OMSSA — OMSSA Search Algorithm - OMSSA web page

☑ Comet — Comet Search Algorithm - Comet web page

☑ Tide — Tide Search Algorithm - Tide web page

☑ Andromeda — Andromeda Search Algorithm - Andromeda web page

Post Processing

□ PeptideShaker - Visualize the results in PeptideShaker

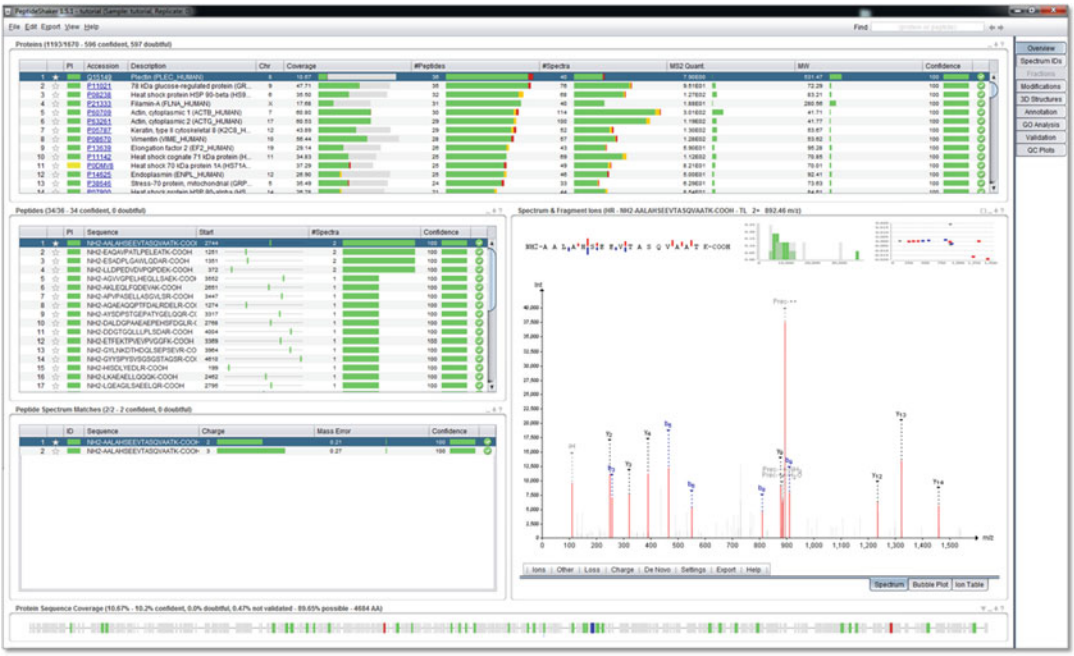
Please cite SearchGUI as Vaudel *et al*.: Proteomics 2011;11(5):996-9.

Fig. 5.1 The main frame of SearchGUI allows the harnessing of multiple search engines. In the *top panel*, the user can select mass spectrometry files, search settings, and an output folder. In the following panels, it is possible to operate msconvert for the conversion and processing of spectra, the different search engines for their identification, and PeptideShaker for the results integration. Note that the cogwheel next to every algorithm can be used to fine tune the settings

be mapped to public resources providing additional information (Vizcaino et al. 2009). It is possible to map proteomic results to genetic information and thus gain knowledge at the genetic level or perform gene ontology (GO) analyses, as done automatically in PeptideShaker when working with UniProt sequences. Peptide sequences and their modifications can also be mapped to the Protein Data Bank (PDB) (Sussman et al. 1998), hence gaining structural information on the post-translational level. Mapping to functional resources like Reactome (Croft et al. 2011) further enables functional analyses. Proteomics data can be combined with the results from other omics technologies such as genomics, transcriptomics, epigenomics, glycomics, metabolomics and lipidomics. Finally, prior to publication, it is required to share all data including the raw files (Martens et al. 2005b; Barsnes and Martens 2013). For this purpose, proteomics data repositories were set up and are coordinated through the ProteomeXchange environment (Vizcaino et al. 2014; Martens et al. 2005a; Craig et al. 2004; Desiere et al. 2006).

5.4 Limitations in Database Searching

While providing an unprecedented throughput with respect to protein identification, database searching also has several limitations, the most



Proteins (1193/1670 - 596 confident, 597 doubtful)

		PI	Accession	Description	Chr	Coverage	#Peptides	#Spectra	MS2 Quant.	MW	Confidence
1	★	🟩	Q15149	Plectin (PLEC_HUMAN)	8	10.87	36	40	7.90E00	531.47	100
2	☆	🟩	P11021	78 kDa glucose-regulated protein (GR...	9	47.71	38	78	9.51E01	72.29	100
3	☆	🟩	P08238	Heat shock protein HSP 90-beta (HS9...	6	35.50	32	68	1.27E02	83.21	100
4	☆	🟩	P21333	Filamin-A (FLNA_HUMAN)	X	17.68	31	40	1.88E01	280.56	100
5	☆	🟩	P60709	Actin, cytoplasmic 1 (ACTB_HUMAN)	7	60.80	30	114	3.01E02	41.71	100
6	☆	🟩	P63261	Actin, cytoplasmic 2 (ACTG_HUMAN)	17	60.53	29	100	1.19E02	41.77	100
7	☆	🟩	P05787	Keratin, type II cytoskeletal 8 (K2C8_H...	12	43.89	29	52	1.30E02	53.67	100
8	☆	🟩	P08670	Vimentin (VIME_HUMAN)	10	56.44	28	57	1.28E02	53.62	100
9	☆	🟩	P13639	Elongation factor 2 (EF2_HUMAN)	19	29.14	26	43	5.90E01	95.28	100
10	☆	🟩	P11142	Heat shock cognate 71 kDa protein (H...	11	34.83	25	69	1.12E02	70.85	100
11	☆	🟨	P0DMV8	Heat shock 70 kDa protein 1A (HS71A...		37.29	25	49	8.21E01	70.01	100
12	☆	🟩	P14625	Endoplasmin (ENPL_HUMAN)	12	26.90	25	48	5.00E01	92.41	100
13	☆	🟩	P38646	Stress-70 protein, mitochondrial (GRP...	5	36.49	24	33	6.29E01	73.63	100

Peptides (34/36 - 34 confident, 0 doubtful)

		PI	Sequence	Start	#Spectra	Confidence
1	★	🟩	NH2-AALAHSEEVTASQVAATK-COOH	2744	2	100
2	☆	🟩	NH2-EAQAVPATLPELEATK-COOH	1281	2	100
3	☆	🟩	NH2-ESADPLGAWLQDAR-COOH	1381	2	100
4	☆	🟩	NH2-LLDPEDVDVPQPDEK-COOH	372	2	100
5	☆	🟩	NH2-AGVVGPELHEQLLSAEK-COOH	3552	1	100
6	☆	🟩	NH2-AKLEQLFQDEVAK-COOH	2651	1	100
7	☆	🟩	NH2-APVPASELLASGVLSR-COOH	3447	1	100
8	☆	🟩	NH2-AQAEAQQPTFDALRDELR-COOH	1274	1	100
9	☆	🟩	NH2-AYSDPSTGEPATYGELQQR-COOH	3317	1	100
10	☆	🟩	NH2-DALDGPAAEAEPEHSFDGLR-COOH	2798	1	100
11	☆	🟩	NH2-DDGTGQLLLPLSDAR-COOH	4004	1	100
12	☆	🟩	NH2-ETFEKTPVEVPVGGFK-COOH	3389	1	100
13	☆	🟩	NH2-GYLNKDTHDQLSEPSEVR-COOH	3964	1	100
14	☆	🟩	NH2-GYYSPYSVSGSGSTAGSR-COOH	4610	1	100
15	☆	🟩	NH2-HISDLYEDLR-COOH	199	1	100
16	☆	🟩	NH2-LKAEAELLQQQK-COOH	2462	1	100
17	☆	🟩	NH2-LQEAGILSAEELQR-COOH	2796	1	100

Peptide Spectrum Matches (2/2 - 2 confident, 0 doubtful)

		ID	Sequence	Charge	Mass Error	Confidence
1	★	🟩	NH2-AALAHSEEVTASQVAATK-COOH	2	0.21	100
2	☆	🟩	NH2-AALAHSEEVTASQVAATK-COOH	3	0.27	100

Spectrum & Fragment Ions (HR - NH2-AALAHSEEVTASQVAATK-COOH - TL 2+ 892.46 m/z)

Protein Sequence Coverage (10.67% - 10.2% confident, 0.0% doubtful, 0.47% not validated - 89.65% possible - 4684 AA)

Fig. 5.2 Upon processing, the results can be mined in the graphical interface of PeptideShaker, as illustrated here with the *Overview* tab, providing a global view on the identified compounds. At the *top*, the protein groups are listed in a table providing details on every group including chromosome mapping when working with UniProt sequences. Clicking on the chromosome number provides more detailed information at the genetic level. To the *left*, the peptides and PSMs for the selected protein group and peptide are listed. The annotated spectrum corresponding to the selected PSM is displayed to the right along with quality control plots. At the *bottom*, a linear representation of the sequence of the leading protein of the selected protein group is displayed with annotated identified peptides and PTMs. The different tabs to the right of the software allow mining the dataset for specific features like PTMs *via* the *Modifications* tab, or structures *via* the *3D Structures* tab

important being that the scope of the search is limited to the content of the database used. To increase the sample coverage by the database, it is possible to extend it to all expected proteins, or to perform so-called error tolerant searches, including isoforms, all types of PTMs, and unanticipated cleavage.

However, such a dramatic increase of the search space leads to very long search time and an increased prevalence of false positives. The false positives can be random matches in the search space or partial matches where the error can be due to a minor detail of the peptide sequence, like the artefactual identification of a PTM or amino acid permutations. The prevalence of both classes of false positives increases with the database size, but while the share of random matches is accurately tracked by the target-decoy strategy (Vaudel et al. 2012a), it is much more challenging to evaluate the rate of non-random false positive matches (Colaert et al. 2011).

Searching very large databases also reduces the distinguishability of the best scoring sequences, and their protein mapping unicity. The post processing of such results will thus be more complex and error prone. As a result, proteomic studies generally restrain the search space by including only the species under consideration, limiting the inclusion of isoforms, of possible cleavage sites, and of variable PTMs. Processing the entire six-frame translation of a complex genome is consequently generally avoided.

At the other end of the scale, it can be tempting to tailor the search space to a list of proteins of interest in order to avoid random matches and reduce processing time. This however results in forcing the algorithm to match specific sequences, and in the absence of competition between the targeted proteins and other possible candidates, the spectra of peptides excluded from the database and resembling target sequences will end up creating non-random false matches, possibly with high scores.

Again, these matches are not well monitored by error rates estimation methods as demonstrated in (Colaert et al. 2011), and the final results will consequently be biased toward the set of targeted sequences. This bias is notably the reason for the systematic inclusion of sequences of contaminant proteins. An illustrative example of the problem that has received worldwide attention is a study on the bee colony collapse disorder (Bromenshenk et al. 2010), where bee samples were searched with a database of potential pathogens, resulting in pathogen identification from spectra more likely to be originating from bee proteins rather than the reporter viruses and fungi (Knudsen and Chalkley 2011).

The quality of a database search thus strongly relies on the subtle balance between covering all proteins present in the sample, and not leaving too much room for false positives. Consequently, low abundant species such as variant of canonical proteins can be missed by the identification procedure, or may fail to pass the validation threshold. One of the most popular methods to alleviate this problem consists in searching with relaxed search settings before filtering. Hence, false positive hits will be diverted outside the scope of the study and subsequently filtered out using target and decoy hits distributions as a guide to identify areas of the search space with high prevalence of true positives (Beausoleil et al. 2006; Vaudel et al. 2011b). This method will yield more peptides if the reduction of the false discovery rate compensates for the loss of hits due to competition in the enlarged search space, and high tolerance search strategies were used to identify novel protein variants (Chick et al. 2015).

Another way to circumvent the problem of large search spaces is to use multiple pass searches, also named iterative searches, where a smaller database of confidently identified proteins is built from the result of a first search, and searched again with more tolerance towards sequence variations and post-translational modifications, a strategy embedded in the X!Tandem search engine (Craig and Beavis 2004). To avoid the demanding preliminary search, targeted databases can also be constructed from public repositories (Shanmugam and Nesvizhskii 2015). Alternatively, the result of the first search can be used to discard spectra of canonical proteins, and the unidentified spectra are then searched for variants (Noble 2015). Databases of increasing complexity can then be searched sequentially. However, the propensity of these methods to introduce non-random false hits, and problems with the applicability of standard error rate estimation techniques make the use of these search strategies subject to debate (Everett et al. 2010; Bern and Kil 2011).

5.5 Application to Proteogenomics

In the trade-off between sample coverage and search space size, proteogenomics provides the opportunity to include additional relevant sequences from the study of the protein expression process. In turn, the identified sequences allow refining the annotation at the genomic and transcriptomic level (Jaffe et al. 2004). Proteogenomics thus interconnects all fields involved in the study of protein expression, genomics, epigenomics, transcriptomics, and proteomics.

Through the availability of next-generation sequencing (NGS) techniques and of high throughput mass spectrometry, it has become affordable to perform multiple omic analyses on the same set of samples. As a result, genome sequencing, RNA sequencing, or ribosome profiling can be used to hypothesize the presence of sample specific variants like specific proteoforms, and protein identification methods can be reciprocally used to confirm their presence.

Multiple applications benefit from such multiomics approaches, as, for example, the study of organisms with little or no genetic annotation, or the investigation of samples containing multiple or no defined species (Muth et al. 2013a). For model organisms like human, a promising application is the use of proteogenomic approaches to investigate the expression of disease specific variants (Alfaro et al. 2014; Zhang et al. 2014), as observed in chemoresistant cancer cell lineages (Pemovska et al. 2013), and from there infer possible effects on the phenotype.

Driven by the need to harness large and heterogeneous datasets from different fields, the need for novel bioinformatic solutions able to conduct multi omics approaches increased. Multiple tools and resources were thus recently established to integrate such information, as reviewed in detail by Menschaert and Fenyö (2015). These include the generation of databases of non-canonical sequences and variants readily usable on proteomics datasets (Sheynkman et al. 2014; Crappe et al. 2015; Risk et al. 2013), as reviewed in details by Nesvizhskii (2014). As a result, standard workflows are being made available able to handle the entire proteogenomic characterization of samples (Boekel et al. 2015).

Due to the high complexity of such tasks, it should be stressed that their use still requires a certain level of bioinformatic expertise. Notably, the amount and complexity of data makes it prohibitive to run these workflows on standard desktop computers, and it is preferable to distribute tasks (Afgan et al. 2012; Trudgian and Mirzaei 2012; Muth et al. 2013b; Verheggen et al. 2014). Proteogenomics workflows can notably be run within Galaxy (Giardine et al. 2005), taking advantage of this powerful environment (Boekel et al. 2015; Fan et al. 2015; Jagtap et al. 2014; Sheynkman et al. 2014).

Proteogenomic approaches also suffer from the fact that non-canonical gene products are low abundant. Since no amplification method is available for amino acid sequences to date, and since low abundant species have a lower probability to trigger mass spectrometry scan events, the discovery potential of new products can be limited by the sensitivity of the proteomic workflow. To increase the chances of finding novel gene products, it is thus necessary to inspect very large amounts of data (Whiteaker et al. 2014).

This can be done by reprocessing publicly available datasets as reviewed in (Vaudel et al. 2016). LNCipedia (Volders et al. 2015) and sORFs.org (Olexiouk et al. 2016) are examples of such strategies, where the abovementioned tools combined with automated search parameter inference (Hulstaert et al. 2013) were applied to mine public ProteomeXchange datasets in proteogenomic contexts using distributed computing (Verheggen et al. 2015). It is however important to note that the mentioned bioinformatic challenges such as local and global error rates estimation are amplified when using these kinds of big data strategies.

5.6 Conclusion

In conclusion, the shortcomings of database search engines and the reliable estimation of error rates are limiting factors of proteogenomics studies. Reducing the dependence on the database, *e.g.*, via better integration of *de novo* sequencing and spectral library searching could be of great help. Spectral libraries can, for example, effectively be used to reduce the number of spectra searched by ruling out commonly observed peptides. In big data strategies, where a large number of experiments are searched for low abundant compounds, it is possible to cluster spectra prior to processing, and mine these clusters of frequently encountered non identified spectra for novel gene products (Griss et al. 2013).

The early years of proteogenomics paved the way for thrilling discoveries, and the dissemination of these approaches will substantially increase the precision of biomedical sample characterization. Currently, proteogenomics studies however require specialized scientific expertise, notably in bioinformatics, impairing the wide application of the approach. The efforts being invested towards the development of standardized, user friendly and interactive workflows with associated training material will certainly help towards overcoming this limitation.

Acknowledgments H.B. and H.R are supported by Bergen Forskningsstiftelse, and H.R. is further supported by Novo Nordisk Fonden and Western Norway Regional

Health Authority. F.B. is supported by the Kristian Gerhard Jebsen foundation.

Conflict of Interest The authors declare no competing financial interests.

References

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature, 422*(6928), 198–207. doi:10.1038/nature01511.

Afgan, E., Chapman, B., & Taylor, J. (2012). CloudMan as a platform for tool, data, and analysis distribution. *BMC Bioinformatics, 13*, 315. doi:10.1186/1471-2105-13-315.

Alfaro, J. A., Sinha, A., Kislinger, T., & Boutros, P. C. (2014). Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature Methods, 11*(11), 1107–1113. doi:10.1038/nmeth.3138.

Altelaar, A. F., Munoz, J., & Heck, A. J. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics, 14*(1), 35–48. doi:10.1038/nrg3356.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L. S. (2004). UniProt: The Universal Protein knowledgebase. *Nucleic Acids Research, 32*(Database issue), D115–D119. doi:10.1093/nar/gkh131.

Barsnes, H., & Martens, L. (2013). Crowdsourcing in proteomics: Public resources lead to better experiments. *Amino Acids, 44*(4), 1129–1137. doi:10.1007/s00726-012-1455-z.

Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., & Gygi, S. P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology, 24*(10), 1285–1292. doi:10.1038/nbt1240.

Bern, M., & Kil, Y. J. (2011). Comment on "Unbiased statistical analysis for multi-stage proteomic search strategies". *Journal of Proteome Research, 10*(4), 2123–2127. doi:10.1021/pr101143m.

Boekel, J., Chilton, J. M., Cooke, I. R., Horvatovich, P. L., Jagtap, P. D., Kall, L., Lehtio, J., Lukasse, P., Moerland, P. D., & Griffin, T. J. (2015). Multi-omic data analysis using Galaxy. *Nature Biotechnology, 33*(2), 137–139. doi:10.1038/nbt.3134.

Bromenshenk, J. J., Henderson, C. B., Wick, C. H., Stanford, M. F., Zulich, A. W., Jabbour, R. E., Deshpande, S. V., McCubbin, P. E., Seccomb, R. A., Welch, P. M., Williams, T., Firth, D. R., Skowronski, E., Lehmann, M. M., Bilimoria, S. L., Gress, J., Wanner, K. W., & Cramer, R. A., Jr. (2010). Iridovirus and microsporidian linked to honey bee colony decline. *PLoS One, 5*(10), e13181. doi:10.1371/journal.pone.0013181.

Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., & Gygi, S. P. (2015). A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology, 33*(7), 743–749. doi:10.1038/nbt.3267.

Colaert, N., Degroeve, S., Helsens, K., & Martens, L. (2011). Analysis of the resolution limitations of peptide identification algorithms. *Journal of Proteome Research, 10*(12), 5555–5561. doi:10.1021/pr200913a.

Coordinators, N. R. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 44*(D1), D7–D19. doi:10.1093/nar/gkv1290.

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology, 26*(12), 1367–1372. doi:10.1038/nbt.1511.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research, 10*(4), 1794–1805. doi:10.1021/pr101065j.

Craig, R., & Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics, 20*(9), 1466–1467. doi:10.1093/bioinformatics/bth092.

Craig, R., Cortens, J. P., & Beavis, R. C. (2004). Open source system for analyzing, validating, and storing protein identification data. *Journal of Proteome Research, 3*(6), 1234–1242. doi:10.1021/pr049882h.

Crappe, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., De Meester, E., De Meyer, T., Van Criekinge, W., Van Damme, P., & Menschaert, G. (2015). PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research, 43*(5), e29. doi:10.1093/nar/gku1283.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., & Stein, L. (2011). Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research, 39*(Database issue), D691–D697. doi:10.1093/nar/gkq1018.

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., & Aebersold, R. (2006). The PeptideAtlas project. *Nucleic Acids Research, 34*(Database issue), D655–D658. doi:10.1093/nar/gkj040.

Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii,

A. I., & Aebersold, R. (2010). A guided tour of the trans-proteomic pipeline. *Proteomics, 10*(6), 1150–1159. doi:10.1002/pmic.200900375.

Diament, B. J., & Noble, W. S. (2011). Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research, 10*(9), 3871–3879. doi:10.1021/pr101196n.

Domon, B., & Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science, 312*(5771), 212–217. doi:10.1126/science.1124619.

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., & Mechtler, K. (2014). MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research, 13*(8), 3679–3684. doi:10.1021/pr500202e.

Elias, J. E., & Gygi, S. P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in Molecular Biology, 604*, 55–71. doi:10.1007/978-1-60761-444-9_5.

Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *Proteomics, 13*(1), 22–24. doi:10.1002/pmic.201200439.

Everett, L. J., Bierl, C., & Master, S. R. (2010). Unbiased statistical analysis for multi-stage proteomic search strategies. *Journal of Proteome Research, 9*(2), 700–707. doi:10.1021/pr900256v.

Fan, J., Saha, S., Barker, G., Heesom, K. J., Ghali, F., Jones, A. R., Matthews, D. A., & Bessant, C. (2015). Galaxy integrated Omics: Web-based standards-compliant workflows for proteomics informed by transcriptomics. *Molecular & Cellular Proteomics, 14*(11), 3087–3093. doi:10.1074/mcp.O115.048777.

French, W. R., Zimmerman, L. J., Schilling, B., Gibson, B. W., Miller, C. A., Townsend, R. R., Sherrod, S. D., Goodwin, C. R., McLean, J. A., & Tabb, D. L. (2015). Wavelet-based peak detection and a new charge inference procedure for MS/MS implemented in ProteoWizard's msConvert. *Journal of Proteome Research, 14*(2), 1299–1307. doi:10.1021/pr500886y.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., & Nekrutenko, A. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research, 15*(10), 1451–1455. doi:10.1101/gr.4086505.

Griss, J., Foster, J. M., Hermjakob, H., & Vizcaino, J. A. (2013). PRIDE Cluster: Building a consensus of proteomics data. *Nature Methods, 10*(2), 95–96. doi:10.1038/nmeth.2343.

Helsens, K., Timmerman, E., Vandekerckhove, J., Gevaert, K., & Martens, L. (2008). Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Molecular & Cellular Proteomics, 7*(12), 2364–2372. doi:10.1074/mcp.M800082-MCP200.

Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J., & Desfeux, A. (2014). OMICtools: An informative directory for multi-omic data analysis. *Database*. doi:10.1093/database/bau069.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., & Clamp, M. (2002). The Ensemble genome database project. *Nucleic Acids Research, 30*(1), 38–41.

Hulstaert, N., Reisinger, F., Rameseder, J., Barsnes, H., Vizcaino, J. A., & Martens, L. (2013). Pride-asap: Automatic fragment ion annotation of identified PRIDE spectra. *Journal of Proteomics, 95*, 89–92. doi:10.1016/j.jprot.2013.04.011.

Jaffe, J. D., Berg, H. C., & Church, G. M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics, 4*(1), 59–77. doi:10.1002/pmic.200300511.

Jagtap, P. D., Johnson, J. E., Onsongo, G., Sadler, F. W., Murray, K., Wang, Y., Shenykman, G. M., Bandhakavi, S., Smith, L. M., & Griffin, T. J. (2014). Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. *Journal of Proteome Research, 13*(12), 5898–5908. doi:10.1021/pr500812t.

Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry, 74*(20), 5383–5392.

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics, 24*(21), 2534–2536. doi:10.1093/bioinformatics/btn323.

Kim, S., & Pevzner, P. A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications, 5*, 5277. doi:10.1038/ncomms6277.

Knudsen, G. M., & Chalkley, R. J. (2011). The effect of using an inappropriate protein database for proteomic data analysis. *PLoS One, 6*(6), e20873. doi:10.1371/journal.pone.0020873.

Kohlbacher, O., Reinert, K., Gropl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., & Sturm, M. (2007). TOPP–the OpenMS proteomics pipeline. *Bioinformatics, 23*(2), e191–e197. doi:10.1093/bioinformatics/btl299.

Lange, E., Gropl, C., Reinert, K., Kohlbacher, O., & Hildebrandt, A. (2006). High-accuracy peak picking of proteomics data using wavelet techniques. In *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pp. 243–254.

Ma, K., Vitek, O., & Nesvizhskii, A. I. (2012). A statistical model-building perspective to identification of MS/

MS spectra with PeptideProphet. *BMC Bioinformatics*, *13*(Suppl 16), S1. doi: 10.1186/1471-2105-13-S16-S1

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., & Apweiler, R. (2005a). PRIDE: The proteomics identifications database. *Proteomics, 5*(13), 3537–3545. doi:10.1002/pmic.200401303.

Martens, L., Nesvizhskii, A. I., Hermjakob, H., Adamski, M., Omenn, G. S., Vandekerckhove, J., & Gevaert, K. (2005b). Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics, 5*(13), 3501–3505. doi:10.1002/pmic.200401302.

Menschaert, G., & Fenyo, D. (2015). Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrometry Reviews*. doi:10.1002/mas.21483.

Muth, T., Benndorf, D., Reichl, U., Rapp, E., & Martens, L. (2013a). Searching for a needle in a stack of needles: Challenges in metaproteomics data analysis. *Molecular BioSystems, 9*(4), 578–585. doi:10.1039/c2mb25415h.

Muth, T., Peters, J., Blackburn, J., Rapp, E., & Martens, L. (2013b). ProteoCloud: A full-featured open source proteomics cloud computing pipeline. *Journal of Proteomics, 88*, 104–108. doi:10.1016/j.jprot.2012.12.026.

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics, 73*(11), 2092–2123. doi:10.1016/j.jprot.2010.08.009.

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods, 11*(11), 1114–1125. doi:10.1038/nmeth.3144.

Nesvizhskii, A. I., & Aebersold, R. (2005). Interpretation of shotgun proteomic data: The protein inference problem. *Molecular & Cellular Proteomics, 4*(10), 1419–1440. doi:10.1074/mcp.R500012-MCP200.

Noble, W. S. (2015). Mass spectrometrists should search only for peptides they care about. *Nature Methods, 12*(7), 605–608. doi:10.1038/nmeth.3450.

Olexiouk, V., Crappe, J., Verbruggen, S., Verhegen, K., Martens, L., & Menschaert, G. (2016). sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research, 44*(D1), D324–D329. doi:10.1093/nar/gkv1175.

Pemovska, T., Kontro, M., Yadav, B., Edgren, H., Eldfors, S., Szwajda, A., Almusa, H., Bespalov, M. M., Ellonen, P., Elonen, E., Gjertsen, B. T., Karjalainen, R., Kulesskiy, E., Lagstrom, S., Lehto, A., Lepisto, M., Lundan, T., Majumder, M. M., Marti, J. M., Mattila, P., Murumagi, A., Mustjoki, S., Palva, A., Parsons, A., Pirttinen, T., Ramet, M. E., Suvela, M., Turunen, L., Vastrik, I., Wolf, M., Knowles, J., Aittokallio, T., Heckman, C. A., Porkka, K., Kallioniemi, O., & Wennerberg, K. (2013). Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discovery, 3*(12), 1416–1429. doi:10.1158/2159-8290.CD-13-0350.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research, 33*(Database issue), D501–D504. doi:10.1093/nar/gki025.

Risk, B. A., Spitzer, W. J., & Giddings, M. C. (2013). Peppy: Proteogenomic search software. *Journal of Proteome Research, 12*(6), 3019–3025. doi:10.1021/pr400208w.

Shanmugam, A. K., & Nesvizhskii, A. I. (2015). Effective leveraging of targeted search spaces for improving peptide identification in Tandem Mass Spectrometry based proteomics. *Journal of Proteome Research, 14*(12), 5169–5178. doi:10.1021/acs.jproteome.5b00504.

Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., Griffin, T. J., & Smith, L. M. (2014). Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics, 15*, 703. doi:10.1186/1471-2164-15-703.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (1998). Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica. Section D: Biological Crystallography, 54*(Pt 6 Pt 1), 1078–1084.

Tabb, D. L., Fernando, C. G., & Chambers, M. C. (2007). MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research, 6*(2), 654–661. doi:10.1021/pr0604054.

Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., & Gojobori, T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Research, 30*(1), 27–30.

Trudgian, D. C., & Mirzaei, H. (2012). Cloud CPFP: A shotgun proteomics data analysis pipeline using cloud and high performance computing. *Journal of Proteome Research, 11*(12), 6282–6290. doi:10.1021/pr300694b.

Vaudel, M., Sickmann, A., & Martens, L. (2010). Peptide and protein quantification: A map of the minefield. *Proteomics, 10*(4), 650–670. doi:10.1002/pmic.200900481.

Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., & Martens, L. (2011a). SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics, 11*(5), 996–999. doi:10.1002/pmic.201000595.

Vaudel, M., Burkhart, J. M., Sickmann, A., Martens, L., & Zahedi, R. P. (2011b). Peptide identification quality control. *Proteomics, 11*(10), 2105–2114. doi:10.1002/pmic.201000704.

Vaudel, M., Burkhart, J. M., Breiter, D., Zahedi, R. P., Sickmann, A., & Martens, L. (2012a). A complex standard for protein identification, designed by evolution. *Journal of Proteome Research, 11*(10), 5065–5071. doi:10.1021/pr300055q.

Vaudel, M., Sickmann, A., & Martens, L. (2012b). Current methods for global proteome identification. *Expert Review of Proteomics, 9*(5), 519–532. doi:10.1586/epr.12.51.

Vaudel, M., Sickmann, A., & Martens, L. (2014a). Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics. *Biochimica et biophysica acta, 1844*(1 Pt A), 12–20. doi:10.1016/j.bbapap.2013.06.019.

Vaudel, M., Venne, A. S., Berven, F. S., Zahedi, R. P., Martens, L., & Barsnes, H. (2014b). Shedding light on black boxes in protein identification. *Proteomics, 14*(9), 1001–1005. doi:10.1002/pmic.201300488.

Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., Martens, L., & Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology, 33*(1), 22–24. doi:10.1038/nbt.3109.

Vaudel, M., Verheggen, K., Csordas, A., Raeder, H., Berven, F. S., Martens, L., Vizcaino, J. A., & Barsnes, H. (2016). Exploring the potential of public proteomics data. *Proteomics, 16*(2), 214–225. doi:10.1002/pmic.201500295.

Verheggen, K., Barsnes, H., & Martens, L. (2014). Distributed computing and data storage in proteomics: Many hands make light work, and a stronger memory. *Proteomics, 14*(4–5), 367–377. doi:10.1002/pmic.201300288.

Verheggen, K., Maddelein, D., Hulstaert, N., Martens, L., Barsnes, H., & Vaudel, M. (2015). Pladipus enables universal distributed computing in proteomics bioinformatics. *Journal of Proteome Research*. doi:10.1021/acs.jproteome.5b00850.

Vizcaino, J. A., Mueller, M., Hermjakob, H., & Martens, L. (2009). Charting online OMICS resources: A navigational chart for clinical researchers. *Proteomics Clinical Applications, 3*(1), 18–29. doi:10.1002/prca.200800082.

Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H.-J., Albar, J. P., Martinez-Bartolome, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., & Hermjakob, H. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology, 32*(3), 223–226. doi:10.1038/nbt.2839. http://www.nature.com/nbt/journal/v32/n3/abs/nbt.2839.html – supplementary-information.

Volders, P. J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J., & Mestdagh, P. (2015). An update on LNCipedia: A database for annotated human lncRNA sequences. *Nucleic Acids Research, 43*(Database issue), D174–D180. doi:10.1093/nar/gku1060.

Whiteaker, J. R., Halusa, G. N., Hoofnagle, A. N., Sharma, V., MacLean, B., Yan, P., Wrobel, J. A., Kennedy, J., Mani, D. R., Zimmerman, L. J., Meyer, M. R., Mesri, M., Rodriguez, H., Clinical Proteomic Tumor Analysis, C., & Paulovich, A. G. (2014). CPTAC assay portal: A repository of targeted proteomic assays. *Nature Methods, 11*(7), 703–704. doi:10.1038/nmeth.3002.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., & Flicek, P. (2016). Ensemble 2016. *Nucleic Acids Research, 44*(D1), D710–D716. doi:10.1093/nar/gkv1157.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J., Liebler, D. C., & Nci, C. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature, 513*(7518), 382–387. doi:10.1038/nature13438.

6 Mutant Proteogenomics

Ákos Végvári

Abstract

Identification of mutant proteins in biological samples is one of the emerging areas of proteogenomics. Despite the fact that only a limited number of studies have been published up to now, it has the potential to recognize novel disease biomarkers that have unique structure and desirably high specificity. Such properties would identify mutant proteoforms related to diseases as optimal drug targets useful for future therapeutic strategies. While mass spectrometry has demonstrated its outstanding analytical power in proteomics, the most frequently applied bottom-up strategy is not suitable for the detection of mutant proteins if only databases with consensus sequences are searched. It is likely that many unassigned tandem mass spectra of tryptic peptides originate from single amino acid variants (SAAVs). To address this problem, a couple of protein databases have been constructed that include canonical and SAAV sequences, allowing for the observation of mutant proteoforms in mass spectral data for the first time. Since the resulting large search space may compromise the probability of identifications, a novel concept was proposed that included identification as well as verification strategies. Together with transcriptome based approaches, targeted proteomics appears to be a suitable method for the verification of initial identifications in databases and can also provide quantitative insights to expression profiles, which often reflect disease progression. Important applications in the field of mutant proteoform identification have already highlighted novel biomarkers in large-scale investigations.

Á. Végvári (✉)
Clinical Protein Science & Imaging, Department of Medical Bioengineering, Biomedical Center, Lund University, Lund, Sweden

Department of Pharmacology & Toxicology, University of Texas Medical Branch, Galveston, TX, USA
e-mail: akos.vegvari@BME.LTH.SE; akvegvar@UTMB.EDU

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_6

Keywords

Non-synonymous single nucleotide polymorphism • Single amino acid variant • Mutation • Bottom-up proteomics • Mass spectrometry • Tandem mass spectra • Protein databases • Mutant protein identification

6.1 Introduction

The technological development in the field of mass spectrometry (MS) based proteomics, in particular bottom-up proteomics, offers a powerful approach to verify newly identified open reading frames (ORFs) and genes (Pandey and Pevzner 2014), which is in line with the Human Genome Organization (HUGO) project. The Human Protein Organization (HUPO) has analogously planned to outline similar goals, including the completion of the human protein catalogue (Paik et al. 2012; Omenn et al. 2015). To facilitate the identification of proteins, databases have been created and maintained, adding and improving protein sequences continuously in these repositories. The most frequently used protein databases (*e.g.*, UniProt/SwissProt and neXtProt) were designed to include the most common forms of human proteins, hence the definition of consensus or canonical protein sequences were established. Due to the interplay between genomic and proteomics research, these protein databases were extended with known splicing isoforms (alternative splicing variants or ASVs), increasing the number of entries by a factor of 2 to 3 (for instance, the neXtProt 2015-09-01 release has 20,066 consensus and 21,932 ASVs). Additionally, great attention has been given to post-translational modifications (PTMs), such as phosphorylation, ubiquitination, glycosylation, etc. Information on PTMs provides a huge amount of novel information about proteoforms with various functions that could be useful in description of disease progression (Nørregaard Jensen 2004). Interestingly, however, little attention has been shown towards mutant proteins, although they represent a level of variability of molecular forms between ASVs and PTMs. In particular, single amino acid variants (SAAVs) are of interest, since:

1. Genetic information/data is often available
2. They may be the ultimate markers as their unique sequence may alter the function
3. They may complicate the quantification of given proteins

The importance of finding non-synonymous single nucleotide polymorphisms (nsSNPs) does not only lie in the discovery of variations in the amino acid sequences that have functional consequences but also to provide information regarding the genetic, and possibly phenotypic, variability within the population of samples (Salisbury et al. 2003). Secondary validation by sequencing of corresponding genomic DNA has confirmed the presence of the predicted single nucleotide polymorphisms (SNPs) in 8 out of 10 SNP-peptides. In their study, Bunger *et al.* highlighted the usefulness of interpreting unassigned spectra as polymorphisms (Bunger et al. 2007). Although, DNA genotyping scans have perhaps the greatest utility in defining the haplotype structure on a genome-wide scale, proteins are a major functional component of most disease progressions. Therefore, information gained from being able to reliably monitor SNP products in proteomic data allows more functional inference to be assigned to expressed alleles. In this regard, the utility of detecting expressed SNPs in proteomics assays will integrate protein profiling with genome information. Such analysis will reveal differential allelic expressions that can be correlated to phenotypic variation between individuals. Recent interest in differential allelic expression has been driven by the discovery that 45–56 % of heterozygous alleles in humans are differentially expressed by a factor of two or more.

6.2 Theoretical Considerations

By definition, a gene mutation is a permanent alteration of the DNA region recognized as a gene. A gene mutation can consist of a number of variations that can be characterized as a single nucleotide variation, a longer sequence changes, all the way up to a large change in a segment of a chromosome that involves multiple genes. On the other hand, from a biological point of view, gene mutations can be divided into two types: hereditary (germline) and acquired (somatic), depending on whether germ or somatic cells are the holders of the mutated DNA. In addition to the most common form of genes (wild type), almost all genes have typical variations due to the high frequency of mutations. If a genetic alteration occurs in more than 1 % of the entire population, such a variation is called a polymorphism. SNPs are frequent in the human genome: as many as 3.1 million SNPs have been found by the International HapMap Consortium (Frazer et al. 2007).

The consequences of a SNP in a gene on the production of the protein can be synonymous, nonsense, and non-synonymous, resulting in unaltered (*i.e.*, fully functional), truncated and mutant protein sequences, respectively (Fig. 6.1). Non-synonymous SNPs certainly contribute to the complexity of the proteome but can also provide significant insight into genetic variability when comparing individuals in a population.

While many of the polymorphisms are harmless and responsible for common variations in humans, some of them can contribute to the risk of disease progression. It is known that somatic mutations can drive cancer development and their accumulation in the mitochondrial DNA is associated with an increased risk of some age-related disorders, such as cardiovascular and neurodegenerative diseases (Taylor and Turnbull 2005). Variations can contribute to an increased likelihood to develop certain diseases on the basis of the genetic makeup of an individual, which is often regarded as genetic predisposition or susceptibility. For instance, we can mention single amino acid variants of the *BRCA1* and *BRCA2* genes that can indicate a significantly increased risk of developing breast and ovarian cancer. *BRCA1* and *BRCA2* are human genes that produce tumor suppressor proteins, which help to repair damaged DNA. Upon mutation their protein products are not made, or do not function correctly, which results in impaired DNA repair, which can in turn lead to an increased probability to develop additional genetic alterations in the cells and eventually cause cancer. The understanding of the underlying biological mechanisms at the molecular level has resulted in an improved clinical diagnosis, monitoring these mutations in risk groups of patients. The level of elevated breast cancer risk may vary as mutations on other genes, like *BARD1* and *BRIP1*, are typically also associated with the disease.

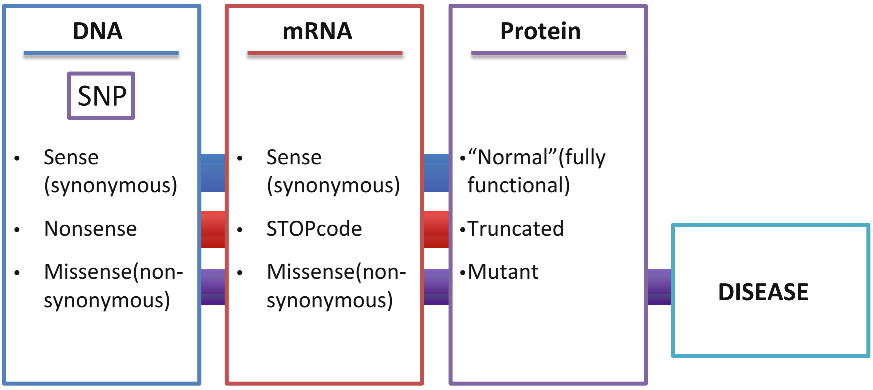


Fig. 6.1 Schematic presentation of single nucleotide polymorphisms in the coding regions of DNA. The non-synonymous SNPs play the most important role in biological mechanisms as their protein products with altered structure and function can contribute to disease progression

Furthermore, protein aggregation has been recognized in neurological disease development, such as Jacob-Krautzfeld disease (associated with prions), Alzheimer's disease and other amyloid diseases (*e.g.*, familial amyloidotic polyneuropathy). A single amino acid alteration may significantly influence protein association and dissociation. For instance, in the case of the familial amyloidotic polyneuropathy (FAP), transthyrein monomer units can aggregate into fibrils that can lead to death upon deposition in the heart and lung. The SAAVs of transthyrein, *p.V30M* and *p.L55P*, facilitate the dissociation of aggregates, while the *p.T119M* proteoform inhibits tetramer dissociation (Hammarström et al. 2003). The main reason for this inhibition is that position 119 is located at the dimer interface. However, low-molecular-weight compounds could be developed to efficiently inhibit transthyrein aggregation by binding to the tetrameric form.

In general, the specific structure of mutant proteoforms allows for the development of efficient drugs. The Pharmaceuticals and Medical Devices Agency in Japan was the first in the world who approved personalized medicine therapy in the case of non-small cell lung cancer, recommending the application of tyrosine kinase inhibitors such as gefitinib and erlotinib. The epidermal growth factor receptor (EGFR) is overexpressed in the cells of certain types of human carcinomas (*e.g.*, in lung and breast cancers), which leads to an inappropriate activation of the anti-apoptotic Ras signaling cascade, eventually causing uncontrolled cell proliferation. It was found that a mutation in the EGFR tyrosine kinase domain is responsible for activating the anti-apoptotic pathways in non-small cell lung cancers (Sordella et al. 2004). These somatic mutations are occur more commonly in lung adenocarcinomas of individuals of Asian descent, women, and non-smokers, rendering erlotinib/gefitinib treatments exceptionally efficient.

It must be noted that the identification of a certain gene as a disease marker is difficult and it is even less probable that this one gene is responsible for disease development alone, indicating that an overall alteration in the genetic profile is expected to be more indicative. Characteristically, variations in individual genes may slightly increase the risk of disease development but the combination of mutations on multiple genes can result in a significant level of risk (Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement 2005). Today, it is generally accepted that the variations in many genes, with their small individual effects, may underlie the susceptibility to many common diseases when combined, this includes diseases like diabetes, obesity, cardiovascular disease and cancer.

Additionally to the understanding of the multifunctional nature of many diseases, and to the tremendously successful genetic investigations on a large number of samples in populations, it is important to emphasize the fact that genes do not have biological function, only their products, the expressed proteins. Therefore, it is necessary to study not only DNA variations but also their corresponding mRNA and their coded proteins, which together can indicate expression activities. Furthermore, the interplay between these principle "omics" areas seems to be inevitable to fully reveal the biology of diseases, unfortunately, the current status of these fields does not always facilitate their integration.

6.3 Methodologies for Detection of Mutant Proteins

6.3.1 Mass Spectrometry Based Proteomics of Mutant Proteoforms

Mass spectrometry based proteomic platforms have the capacity to identify a great number of proteins simultaneously in a single analysis. However, protein identifications by the most commonly applied bottom-up proteomics approach largely rely on protein sequence databases. Due to the fact that such databases are comprised of consensus sequences, *i.e.*, the most frequently observed proteoforms, mass spectra of SAAVs and many other modifications of the analyzed peptides will be unassigned. In principle, known ASVs can be identified searching databases, like neXtProt and UniProt, if isoform specific tryptic peptides could be generated prior to MS analysis. Alternative digestion strategies, using a combination of proteases for the generation of specific peptide sequences may improve ASV identifications. However, the identifications of SAAVs in high quality MS/MS data sets can be easily achieved developing searchable databases that include altered amino acid sequences (Nesvizhskii et al. 2006). Genomic information of nsSNPs can be translated and incorporated into protein sequence databases that can confirm genomic based data in existing tandem mass spectra, and eventually observe novel proteoforms in biological samples.

6.3.1.1 Databases for Identification of SAAVs

The first reported database designed for mutant protein identification was based on the International Protein Sequence (IPI) database that was widely used, holding some 70,000 human protein entries (Schandorff et al. 2007). The inclusion of SAAVs and single amino acid conflicts reported in the SwissProt databases posed the problem of increased sequence redundancy that could eventually compromise the confidence of identifications. Therefore, sequential variations of proteins were attached to the consensus protein sequences with an addition of letter “J” (as a “spacer” to recognize the extra information in entries) between each peptide with the mutation site flanked with a tryptic peptide at both ends. The resulting MSIPI database was completed with a decoy database and published together with each new IPI release by EBI until its final version (v3.67) that held 87,062 human protein entries.

Alternatively, another protein database (K-SNPdb) was created with 125,622 tryptic peptides, which included sequences of the altered amino acids (Bunger et al. 2007). The construction was based on filtering the NCBI dbSNP for nsSNPs (Sherry et al. 2001) that exceeded ten million SNPs throughout the entire human genome. The number of nsSNPs out of all coding region SNPs was about 65,000. In addition to the filtering, manual allocation with the protein accession numbers, the location of nsSNPs and the amino acid changes were derived from the NCBI protein database, creating a fasta file with paired reference and alternative alleles. In order to improve the false discovery rate, a decoy database (FalseSNPdb) with the same number of peptides and identical masses to the peptides of K-SNPdb was composed from IPI entries. This strategy granted the identification of 629 SAAVs, of which 36 were not present in the protein databases of NCBI and IPI.

In an attempt to collect comprehensive sequential data about human mutant proteins, in particular about those involved in cancer, oncogenesis and tumor progression, a novel database (CanProVar) was created (Li et al. 2010). Information on protein variations from public resources, including the Human Proteome Initiative (HPI) (O’Donovan et al. 2001), the Catalogue of Somatic Mutations in Cancer (COSMIC) (Bamford et al. 2004), the Online Mendelian Inheritance in Man (OMIM) (Hamosh et al. 2005), the Cancer Genome Atlas (TCGA) (Comprehensive genomic characterization defines human glioblastoma genes and core pathways 2008) and two large-scale cancer genome sequencing studies (Greenman et al. 2007; Sjoblom et al. 2006), were integrated with a special recognition of cancer related variations (crVAR). The final version of CanProVar holds 41,541 non-cancer specific and 11,445 cancer related variations (http://bioinfo.vanderbuilt.edu/canprovar/). Most importantly, this collection of human mutant proteins is searchable on-line, offering extremely useful data linked to cancer samples, additional data sources, publications along with functional information on gene ontology annotations and interaction partners (Fig. 6.2).

As an outcome of the CanProVar project, an effective identification workflow with multiple search engine options and a tool designed for the correction of the false discovery rate (FDR), was proposed and demonstrated using CRC cell line data (Li et al. 2011). Notably, this downloadable MS-CanProVar database was completed with the Ensemble protein database (v53), 148

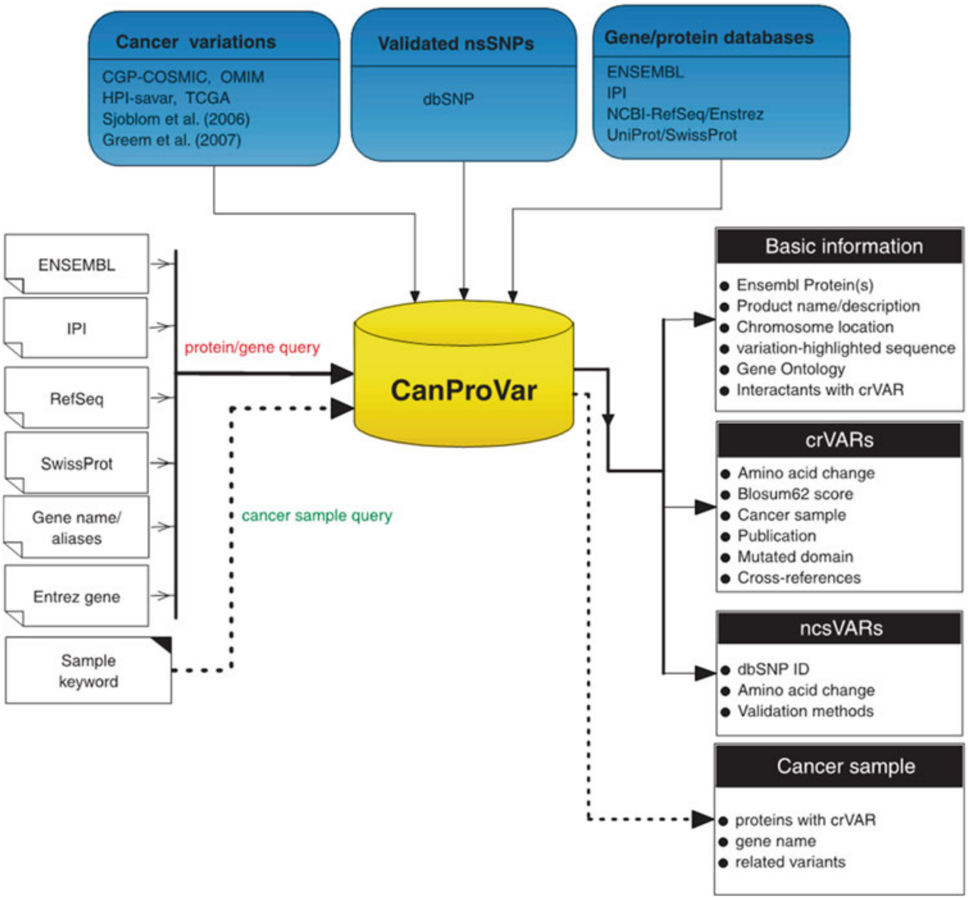


Fig. 6.2 The system architecture of CanProVar protein database that provides a large number of SAAV sequences (Reproduced with permission from Li et al. 2010) Copyright (2010) John Wiley and Sons

contaminant sequences and their reversed sequences as decoys (total searchable entry number is 290,440). Missense variations, non-sense variations and single amino acid deletions and insertions were all included in the database. To address the increasing redundancy posed by the inclusion of mutant sequences, they were shortened to a tryptic peptide with the mutation site flanked by two peptides. As a result, the addition of these new peptides increased the databased size by only 3.4 %.

The MS-CanProVar database formed the basis of a recent approach that combined it with the unique variant sequences in UniProt human polymorphisms (release 2011-12-14), resulting in a total of 87,745 SAAVs (Song et al. 2014). The Swiss-CanSAAVs database contains a total of 161,747 downloadable entries (http://bioanalysis.dicp.ac.cn/proteomics/Publications/SSAV/SAAV-Database.htm) with minimized redundancy using only tryptic mutant peptide sequences flanked by two additional peptides. A customized database, the Human Protein Mutant Database (HPMD), was also created by extracting and combining sequential information of known disease mutations from OMIM, the Protein Mutant Database (PMD) (Kawabata et al. 1999), the Systematic Platform for Identifying Mutated Proteins (SysPIMP) (Xi et al. 2009) and UniProt (Magrane and Consortium 2011) (see Fig. 6.3) (Mathivanan et al. 2012). To improve the FDR, sequence redundancy was decreased by limiting the peptide sequences to 101 amino acids with the mutation site in the center position

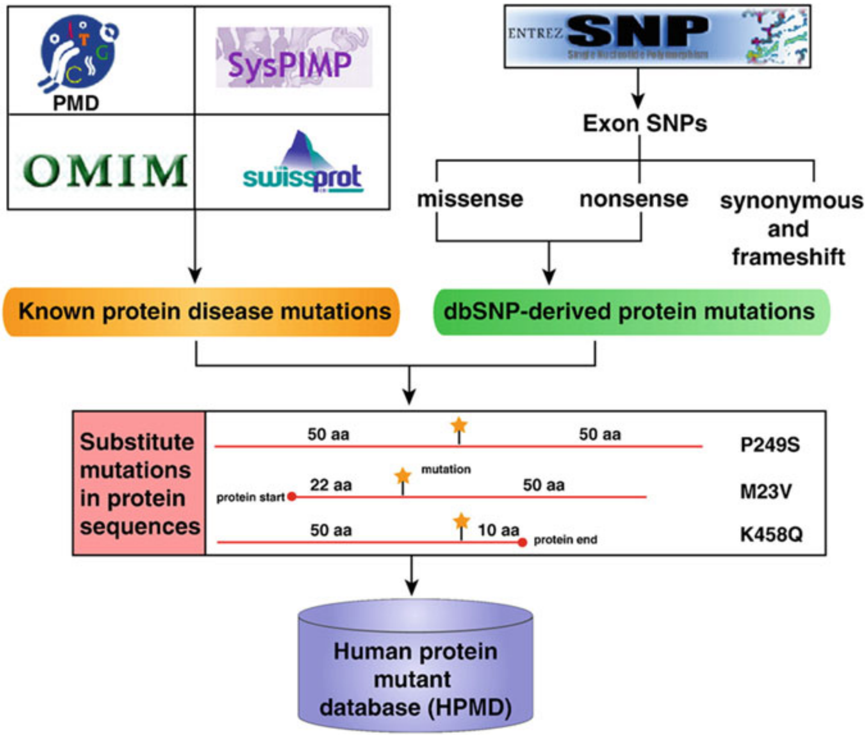

PMD
SysPIMP
OMIM
swissprot
ENTREZ SNP
Exon SNPs
missense
nonsense
synonymous and frameshift
Known protein disease mutations
dbSNP-derived protein mutations
Substitute mutations in protein sequences
50 aa
50 aa
P249S
protein start
22 aa
mutation
50 aa
M23V
50 aa
10 aa
protein end
K458Q
Human protein mutant database (HPMD)

Fig. 6.3 Construction of Human Protein Mutant Database (HPMD) for MS based protein mutation search. Schematic of the construction of HPMD is shown. Known protein disease mutations downloaded from OMIM, PMD, SysPIMP and UniProt were combined along with the missense and nsSNPs from dbSNP. The mutations were substituted in protein sequences to form peptides (maximum 101 amino acids) with mutations. The mutations were fixed to the center (51st residue) unless and until the mutation is localized close to the start or end of the protein sequence. The database composed 171,919 mutations (31,479 – known disease mutations and 140,440 – dbSNP) (Reproduced with permission from Mathivanan et al. 2012) Copyright (2012) Elsevier

(51st residue). This strategy has a drawback as it excludes the possible identification of mutant proteins by a miss-cleaved tryptic peptide.

Through the increasing access to RNA-Seq data, another path has opened to generate extended protein sequence databases, which is the translation of transcriptomic information to amino acids. The combination of shotgun proteomics with next generation sequencing (NGS) technologies has shown to be an effective approach to gain multilevel information and knowledge about cellular systems (Chen et al. 2012). To facilitate translations of RNA-Seq sequences to protein levels, an elegant bioinformatics tool was introduced recently allowing for the generation of customized protein databases (Wang and Zhang 2013). The R package of customProDB can easily create improved protein databases from RNA-Seq data with identified single nucleotide variations, short insertions and deletions as well as novel junctions between exons. The customProDB was an integrated and important part of the newly developed proteogenomic dashboard (dasHPPboard) intended to facilitate the protein mapping efforts of HPP (Tabas-Madrid et al. 2015).

6.3.2 Concept for Identification of Mutant Proteins

The strategy to identify novel mutant proteoforms in biological samples was designed using high quality shotgun proteomic tandem mass spectra for database search by existing search engine algorithms (Lichti et al. 2015). The key element of the approach was the unique set of database entries that describes all SAAV sequences, translated from known genomic studies (Ensembl). Using a custom made software tool, new protein sequences were generated to include a point mutation in each new entry that thus differed in a single amino acid from the consensus protein. The mutant protein database (MuPdb) included 2.3 million SAAVs, excluding titin (Q8WZ42). The sequence redundancy was greatly reduced keeping the tryptic peptides with the mutation site surrounded by two missed cleavages at both termini. The resulting *in silico* derived proteoforms were denoted following the neXtProt nomenclature, including the access codes but also adding information about the nature and the position of the mutation (such as NX_P07288-SNP-L-132-I).

The MuPdb was rendered as a combination of consensus (40,548 entries of UniProtKB) and mutant proteoform sequences of chromosome 19 (132,264 entries), together with 115 common contaminant sequences (cRAP) in standardized. *fasta* format, in order to be used with various search engines, including Proteome Discoverer, Mascot and PEAKS. To address the challenge that the large search space represents, a custom decoy database was created using PEAKS. Additionally, manual validation of tandem mass spectra was performed following blast searches for uniqueness. An initial identification and validation on glioblastoma stem cells (GSC) revealed many SAAVs. Interestingly, a thoroughly investigated mutation (*p.T186R*) of branched-chain aminotransferase 2 (BCAT2) was confirmed (Lichti et al. 2015). This and other newly observed SAAVs in GSC samples were further validated at the transcript level and by SRM-assays designed for suitable SAAV peptides.

This concept was generalized for the identification of SAAVs in any biological sample as presented in Fig. 6.4. Following searches of high quality MS/MS data in the custom made mutant database, the initial findings need to be verified. Currently, this step consists of both targeted proteomics and transcriptomic methods, a combination of which is sufficiently powerful to provide novel biomarkers and drug targets in future applications.

The SRM-MS analysis for verification of mutant proteoforms, targeting the most potential tryptic peptides specific to mutation sites and their corresponding wild type sequences, can also be performed. Synthetic heavy isotope labeled peptides with corresponding sequences can be spiked into the clinical samples for unambiguous identification of mutant proteoforms. In addition, to provide qualitative confirmation and quantification, the ratios between wild type and mutant forms can be determined in heterozygous expression.

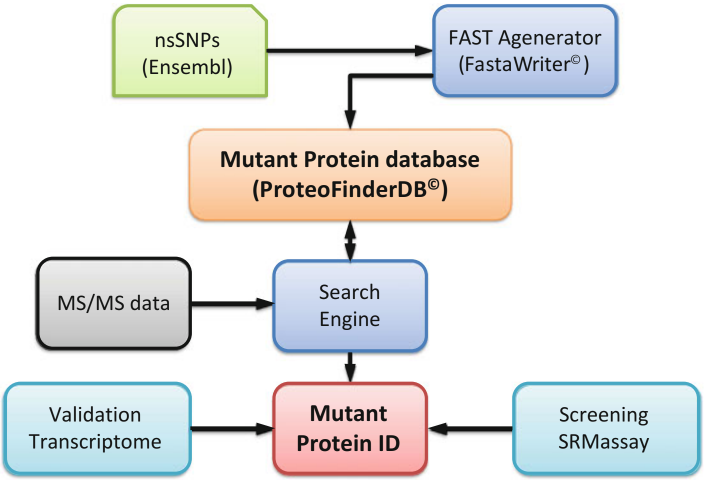
Furthermore, RNA sequence analysis can be performed with biological samples and the transcript data can be used to verify newly identified mutant proteins. This approach can provide considerable reference information and verify SAAVs. The quantitative readout of RNA-Seq results can also be correlated with observed levels of expressed mutant proteins for confirmation.

6.3.3 Targeted Proteomics of SAAVs

One of the most popular mass spectrometry technologies, selected reaction monitoring (SRM), can be successfully applied to identify and quantify specific peptides within the digested samples of complex mixtures (Feng and Picotti 2016). In addition, the SRM methodology is inherently easy to multiplex, allowing for the development of multiple protein assays that offer high sensitivity and throughput. When applying the stable isotope technology, uniformly ^{13}C-^{15}N-labeled proteins can be quantified in blood plasma at levels of 100 ng/mL. However, in many cases, additional enrichment steps are required for the identification of proteins present at lower concentrations in human samples like plasma or serum. Targeted enrichment, with or without antibodies, has been introduced 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; Whiteaker et al. 2011), and mass spectro-

Fig. 6.4 The general concept of identification of SAAVs by tandem mass spectra searching a specialized protein database containing more than two million SAAV sequences



metric 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). Immunoprecipitation (IP) is a logical strategy to enrich mutant proteoforms in combination with targeted proteomics techniques, such as SRM, because antibodies can be generated against most proteins of interest and SRM-MS does not require absolute specificity for the antigens or for the mutations of interest. Additionally, IP can remove the most abundant proteins, including cytoskeletal proteins, immunoglobulins, and serum albumin from biological samples (Anderson et al. 2004).

Since antibodies are not always available and can be 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).

6.3.4 Quantifications of SAAVs

The SRM technology offers precise and efficient quantifications of known proteins, targeting their characterized proteotypic peptides in biological samples. Utilizing the high specificity of SRM-MS, multiplexing can be easily achieved, providing a relatively high throughput methodology. However, the requirement of isotope labeled peptide standards for relative or absolute quantification may constrain this approach, considering the difficulties to synthesize certain peptide sequences, the related costs may also be a limitation.

Nevertheless, quantitative analysis of mutant proteoforms in studies of disease progression can certainly provide important insights into the ratio of allele-specific gene expressions, which has been shown to be closely associated with variations in individuals (Yan et al. 2002; Montgomery et al. 2010). It has been also demonstrated that a SAAV of a single allele could in fact be expressed in either a homozygous or heterozygous manner (Végvári et al. 2013). The frequency of the mutant prostate specific antigen (PSA_*p.L132I*) form agreed well with population based genomic data, although more systematic and large-scale studies are required to understand how universal this finding is (Fig. 6.5). There are indications that disease progression may be monitored by the level of mutant proteoforms in heterozygous expressions, which can improve our understanding of, for instance, cancer biology.

An encouraging investigation was designed to utilize SRM based targeted proteomics for detection and quantification of selected SAAVs in 290 clinical plasma samples collected from Asian patients with both obesity and diabetes (Su et al.

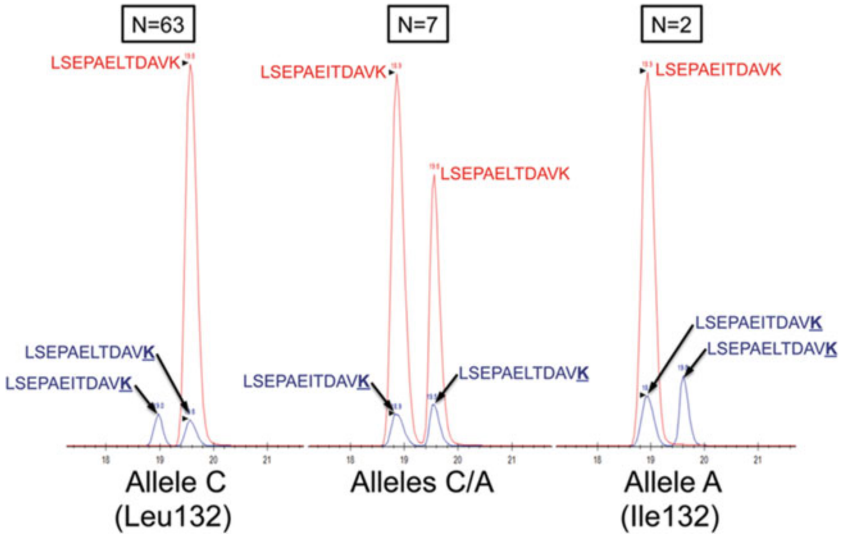


Fig. 6.5 Detection of three possible combinations of allele expressions in examples of SRM-MS analyses in clinical samples. Endogenous signals of LSEPAELTDAVK and LSEPAEITDAVK are shown in *red*, and their corresponding heavy-isotope labeled internal standard signals are in *blue* (Reproduced with permission from Végvári et al. 2013)

2011). Following initial identifications of SAAVs in patient groups, including healthy controls, key proteotypic tryptic peptides were monitored and their corresponding complement component proteins (C7, factor H and C5) were determined by absolute quantification. The results indicated that the homozygous expressions of wild type C7_*p.T587* and factor H_*p.V62* were over represented in the selected Asian patient groups, while the C7_*p.T587P* was significantly higher in control samples. Additionally, no homozygous expression was found in any individuals, which agreed well with previous studies (Gimelbrant et al. 2007). Similarly, heterozygous expression profiles of these SAAVs were determined with sufficient accuracy. This study has proven the novel concept of SRM-based quantitative analysis, which indicated that the levels of heterozygous and homozygous SAAVs in patient populations have significant associations with certain disease traits.

A cost effective quantification approach was developed for the large-scale study of SAAVs in clinical samples (Song et al. 2014). The stable isotope dimethyl-labeling methodology (Kovanich et al. 2012) could be adopted to quantify a total of 282 unique SAAV peptides in combined CID and HCD tandem mass spectra (Fig. 6.6). Because leucine (Leu) and isoleucine (Ile) are isobaric, SAAVs with altered Leu or Ile were excluded in the final results. The initial identification of SAAVs was performed using searches of mass spectra against a custom made protein database, holding 87,745 amino acid variant sequences and 73,910 UniProt canonical protein entries (Swiss-CanSAAVs). Notably, the mutant sequences were shortened to a tryptic peptide with two missed cleavage at both ends in order to reduce sequential redundancy and thus improve the false discovery rate (FDR). Furthermore, the Swiss-CanSAAVs database was concatenated with the reversed sequences allowing for FDR analysis.

6.4 Applications and Their Biological Findings

Up to today, little attention has been paid to the functional link between mutant proteins and diseases (Wang et al. 2011). Cancer research has recently found that solid tumors typically produce 20–100 mutant genes with non-synonymous

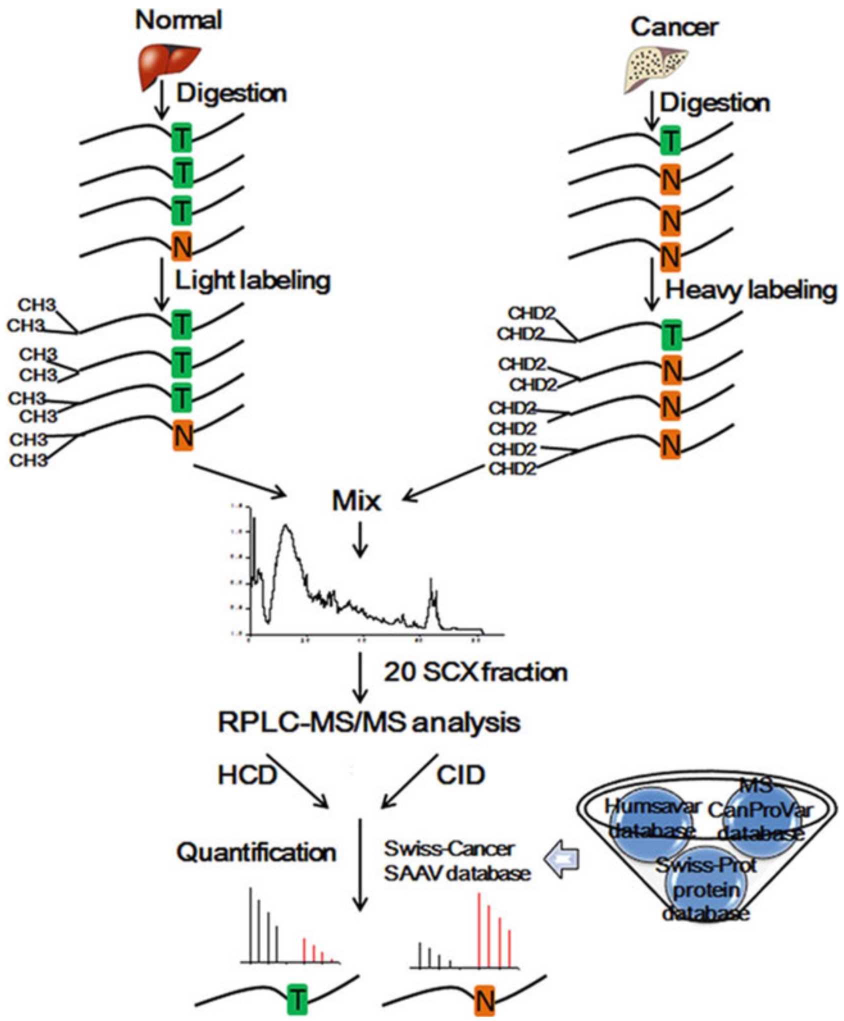


Fig. 6.6 Workflow for the large-scale quantitative analysis of SAAVs between hepatocellular carcinoma and normal human liver tissues (Reproduced with permission from Song et al. 2014. Copyright (2014) American Chemical Society)

alterations, as DNA-based sequencing studies have revealed (Wood et al. 2007). Recent studies on certain cancer forms have shown that mutant proteins can be associated with disease and even be the cause of disease (Bozic et al. 2010; Haber and Settleman 2007). Based on the occurrence and biological function of these mutant proteins, two classes were suggested:

1. “Drivers” that can initiate and are responsible for tumor genesis
2. “Passengers” are not directly associated with malignant differentiation (Haber and Settleman 2007; Bozic et al. 2010).

Importantly, the altered genetic codes of nsSNPs were found to be tightly associated with physiological and pathological traits of individuals (Sun et al. 2008). Additionally, the ratio of allele-specific gene expressions in heterozygous state is also associated with various traits of individuals (Montgomery et al. 2010) and the quantitative relationship of the wild type/mutant proteins can also indicate disease traits (Yan et al. 2002; Montgomery et al. 2010). Consequently, both qualitative and quantitative data about the structures and the functional proteins of individuals with SAAVs are required for comprehensive analysis.

Identification and quantification of mutant proteoforms were performed searching tandem mass spectra against a custom database that consisted of 87,745 amino acid variant sequences and all canonical protein entries of UniProtKB (Song et al. 2014). The approach was applied on profiling mutant proteomes in hepatocellular carcinoma (HCC) and healthy human tissue samples identifying 282 unique SAAV sites. Importantly, a significant increase of carbamoyl phosphate synthase (CPS1) *p.T1406N* and HIV-1 TAT-interactive protein 2 (HTATIP2) *p.S197R* mutations were quantified in HCC samples, which could be associated with cancer progression (Song et al. 2014). Similarly significant alteration of mutant proteomes was detected in serum samples from patients with pancreatic cancer and quantified using a isobaric labeling method (Nie et al. 2014). As a result, a novel biomarker panel was suggested, including α-1-antichymotrypsin (AACT), thrombospondin-1 (THBS1) and a mutant form of serotransferrin (TF_*p.V448I*), that could differentiate pancreatic cancer from healthy controls and chronic pancreatitis.

6.5 Future Perspective

Many genomic studies have produced a large amount of high quality data originating from population wide investigations. While the association of genes with certain diseases, identifying germline and somatic mutations, is very useful, the actual expression profiles of their wild type and mutant products is at least as important, given that proteins are the functional players in biology. Because the altered biology of cells, characteristic of disordered progresses, is driven by proteins, functional data should be generated by taking snapshots of expression profiles in healthy and patient samples. MS-based proteomics provides a unique tool to assess the expression profiles of mutant proteins in body fluids and tissue samples, which can be identified as lead candidates of optimal disease biomarkers. The qualitative and quantitative analyses of these proteoforms could thus result in novel diagnostic and prognostic values.

The fact that SAAVs can be identified in tandem mass data by unique peptide sequences, which are absent from typical protein databases, makes their observation difficult to confirm. Theoretically, certain SAAV and also ASV specific peptides may be identical with tryptic sequences of other consensus proteins. Additionally, mutant proteins can have altered biological activities, making these proteoforms functionally new molecules. It may be suggested that the definition of proteins might be improved with a more functional view. Of course, as the identification is strictly based on structural information, such a new protein definition would not be directly supportive if databases are confined to consensus sequences only. However, a more progressive view and protein definition should facilitate the development of identification strategies to identify mutant proteins in large-scale clinical studies.

References

Anderson, N. L., Anderson, N. G., Haines, L. R., Hardie, D. B., Olafson, R. W., & Pearson, T. W. (2004). Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *Journal of Proteome Research, 3*(2), 235–244.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., & Wooster, R. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer, 91*(2), 355–358.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., & Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences, 107*(43), 18545–18550.

Bunger, M. K., Cargile, B. J., Sevinsky, J. R., Deyanova, E., Na, Y., Hendrickson, R. C., & Stephenson, J. L. (2007). Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *Journal of Proteome Research, 6*, 2331–2340.

Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y. K., Chen, R., Miriami, E., Karczewski, K. J., Hariharan, M., Dewey, F. E., Cheng, Y., Clark, M. J., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J. T., Hillenmeyer, S., Haraksingh, R., Sharon, D., Euskirchen, G., Lacroute, P., Bettinger, K., Boyle, A. P., Kasowski, M., Grubert, F., Seki, S., Garcia, M., Whirl-Carrillo, M., Gallardo, M., Blasco, M. A., Greenberg, P. L., Snyder, P., Klein, T. E., Altman, R. B., Butte, A. J., Ashley, E. A., Gerstein, M., Nadeau, K. C., Tang, H., & Snyder, M. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell, 148*(6), 1293–1307.

Comprehensive genomic characterization defines human glioblastoma genes and core pathways (2008). *Nature, 455*(7216), 1061–1068.

Feng, Y., & Picotti, P. (2016). Selected reaction monitoring to measure proteins of interest in complex samples: A practical guide. *Methods in Molecular Biology (Clifton, NJ), 1394*, 43–56.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449 (7164):851–861.

Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: Recommendation statement. (2005). Annals of internal medicine 143 (5):355–361

Gimelbrant, A., Hutchinson, J. N., Thompson, B. R., & Chess, A. (2007). Widespread monoallelic expression on human autosomes. *Science, 318*(5853), 1136–1140.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y. E., DeFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M. H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., & Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*(7132), 153–158.

Haber, D. A., & Settleman, J. (2007). Cancer: Drivers and passengers. *Nature, 446*(7132), 145–146.

Hammarström, P., Wiseman, R. L., Powers, E. T., & Kelly, J. W. (2003). Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science (New York, NY), 299*(5607), 713–716.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research, 33*(Database issue), D514–517.

Kawabata, T., Ota, M., & Nishikawa, K. (1999). The protein mutant database. *Nucleic Acids Research, 27*(1), 355–357.

Kovanich, D., Cappadona, S., Raijmakers, R., Mohammed, S., Scholten, A., & Heck, A. J. (2012). Applications of stable isotope dimethyl labeling in quantitative proteomics. *Analytical and Bioanalytical Chemistry, 404*(4), 991–1009.

Li, J., Duncan, D. T., & Zhang, B. (2010). CanProVar: A human cancer proteome variation database. *Human Mutation, 31*(3), 219–228.

Li, J., Su, Z., Ma, Z. Q., Slebos, R. J. C., Halvey, P., Tabb, D. L., Liebler, D. C., Pao, W., & Zhang, B. (2011). A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Molecular & Cellular Proteomics, 10*(5), M110.006536–M006110.006536.

Lichti, C. F., Mostovenko, E., Wadsworth, P. A., Lynch, G. C., Pettitt, B. M., Sulman, E. P., Wang, Q., Lang, F. F., Rezeli, M., Marko-Varga, G., Végvári, Á., & Nilsson, C. L. (2015). Systematic identification of single amino acid variants in Glioma stem-cell-derived chromosome 19 proteins. *Journal of Proteome Research, 14*(2), 778–786.

Lopez, M. F., Rezai, T., Sarracino, D. A., Prakash, A., Krastins, B., Athanas, M., Singh, R. J., Barnidge, D. R., Oran, P., Borges, C., & Nelson, R. W. (2010). Selected reaction monitoring-mass spectrometric immunoassay responsive to parathyroid hormone and related variants. *Clinical Chemistry, 56*(2), 281–290.

Magrane, M., & Consortium, U. (2011). UniProt knowledgebase: A hub of integrated protein data. *Database* 2011, bar009.

Mathivanan, S., Ji, H., Tauro, B. J., Chen, Y.-S, & Simpson, R. J. (2012). Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *Journal of Proteomics, 76*, 141–149.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., & Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature, 464*(7289), 773–777.

Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Gruissem, W., Baginsky, S., & Aebersold, R. (2006). Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Molecular & Cellular Proteomics, 5*(4), 652–670.

Nie, S., Yin, H., Tan, Z., Anderson, M. A., Ruffin, M. T., Simeone, D. M., & Lubman, D. M. (2014). Quantitative analysis of single amino acid variant peptides associated with pancreatic cancer in serum by an isobaric labeling quantitative method. *Journal of Proteome Research, 13*(12), 6058–6066.

Nørregaard Jensen, O. (2004). Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology, 8*(1), 33–41.

O'Donovan, C., Apweiler, R., & Bairoch, A. (2001). The human proteomics initiative (HPI). *Trends in Biotechnology, 19*(5), 178–181.

Omenn, G. S., Lane, L., Lundberg, E. K., Beavis, R. C., Nesvizhskii, A. I., & Deutsch, E. W. (2015). Metrics for the Human Proteome Project 2015: Progress on the human proteome and guidelines for high-confidence protein identification. *Journal of Proteome Research, 14*(9), 3452–3460.

Paik, Y. K., Omenn, G. S., 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, M. S., Yoo, J. S., Garin, J., Archakov, A., Bergeron, J., Salekdeh, G. H., & Hancock, W. S. (2012). Standard guidelines for the chromosome-centric human proteome project. *Journal of Proteome Research, 11*(4), 2005–2013.

Pandey, A., & Pevzner, P. A. (2014). Proteogenomics. *Proteomics, 14*(23–24), 2631–2632.

Salisbury, B. A., Pungliya, M., Choi, J. Y., Jiang, R., Sun, X. J., & Stephens, J. C. (2003). SNP and haplotype variation in the human genome. *Mutation Research, 526*, 53–61.

Schandorff, S., Olsen, J. V., Bunkenborg, J., Blagoev, B., Zhang, Y., Andersen, J. S., & Mann, M. (2007). A mass spectrometry-friendly database for cSNP identification. *Nature Methods, 4*, 465–466.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research, 29*, 308–311.

Shi, T., Fillmore, T. L., Sun, X., Zhao, R., Schepmoes, A. A., Hossain, M., Xie, F., Wu, S., Kim, J. S., Jones, N., Moore, R. J., Pasa-Tolic, L., Kagan, J., Rodland, K. D., Liu, T., Tang, K., Camp, D. G., 2nd, Smith, R. D., & Qian, W. J. (2012). Antibody-free, targeted mass-spectrometric approach for quantification of proteins at low picogram per milliliter levels in human plasma/serum. *Proceedings of the National Academy of Sciences of the United States of America, 109*(38), 15395–15400.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., & Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science, 314*(5797), 268–274.

Song, C., Wang, F., Cheng, K., Wei, X., Bian, Y., Wang, K., Tan, Y., Wang, H., Ye, M., & Zou, H. (2014). Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. *Journal of Proteome Research, 13*(1), 241–248.

Sordella, R., Bell, D. W., Haber, D. A., & Settleman, J. (2004). Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. *Science, 305*(5687), 1163–1167.

Su, Z. D., Sun, L., Yu, D. X., Li, R. X., Li, H. X., Yu, Z. J., Sheng, Q. H., Lin, X., Zeng, R., & Wu, J. R. (2011). Quantitative detection of single amino acid polymorphisms by targeted proteomics. *Journal of Molecular Cell Biology, 3*(5), 309–315.

Sun, T., Zhou, Y., Yang, M., Hu, Z., Tan, W., Han, X., Shi, Y., Yao, J., Guo, Y., Yu, D., Tian, T., Zhou, X., Shen, H., & Lin, D. (2008). Functional genetic variations in cytotoxic T-lymphocyte antigen 4 and susceptibility to multiple types of cancer. *Cancer Research, 68*(17), 7025–7034.

Tabas-Madrid, D., Alves-Cruzeiro, J., Segura, V., Guruceaga, E., Vialas, V., Prieto, G., Garcia, C., Corrales, F. J., Albar, J. P., & Pascual-Montano, A. (2015). Proteogenomics dashboard for the human proteome project. *Journal of Proteome Research, 14*(9), 3738–3749.

Taylor, R. W., & Turnbull, D. M. (2005). Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics, 6*(5), 389–402.

Végvári, Á., Sjödin, K., Rezeli, M., Malm, J., Lilja, H., Laurell, T., & Marko-Varga, G. (2013). Identification of a novel proteoform of prostate specific antigen (SNP-L132I) in clinical samples by multiple reaction monitoring. *Molecular & Cellular Proteomics, 12*(10), 2761–2773.

Wang, X., & Zhang, B. (2013). CustomProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics (Oxford, UK), 29*(24), 3235–3237.

Wang, Q., Chaerkady, R., Wu, J., Hwang, H. J., Papadopoulos, N., Kopelovich, L., Maitra, A., Matthaei, H., Eshleman, J. R., Hruban, R. H., Kinzler, K. W., Pandey, A., & Vogelstein, B. (2011). Mutant proteins as cancer-specific biomarkers. *Proceedings of the National Academy of Sciences, 108*(6), 2444–2449.

Whiteaker, J. R., Zhao, L., Zhang, H. Y., Feng, L. C., Piening, B. D., Anderson, L., & Paulovich, A. G. (2007). Antibody-based enrichment of peptides on magnetic beads for mass-spectrometry-based quantification of serum biomarkers. *Analytical Biochemistry, 362*(1), 44–54.

Whiteaker, J. R., Zhao, L., Abbatiello, S. E., Burgess, M., Kuhn, E., Lin, C., Pope, M. E., Razavi, M., Anderson, N. L., Pearson, T. W., Carr, S. A., & Paulovich, A. G. (2011). Evaluation of large scale quantitative proteomic assay development using peptide affinity-based mass spectrometry. *Molecular & Cellular Proteomics, 10*(4), M110.005645.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyankasky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., & Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science (New York, NY), 318*(5853), 1108–1113.

Xi, H., Park, J., Ding, G., Lee, Y. H., & Li, Y. (2009). SysPIMP: The web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry. *Nucleic Acids Research, 37*(Database issue), D913–920.

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science, 297*(5584), 1143.

7 Proteogenomic Analysis of Single Amino Acid Polymorphisms in Cancer Research

Alba Garin-Muga, Fernando J. Corrales,
and Victor Segura

Abstract

The integration of genomics and proteomics has led to the emergence of proteogenomics, a field of research successfully applied to the characterization of cancer samples. The diagnosis, prognosis and response to therapy of cancer patients will largely benefit from the identification of mutations present in their genome. The current state of the art of high throughput experiments for genome-wide detection of somatic mutations in cancer samples has allowed the development of projects such as the TCGA, in which hundreds of cancer genomes have been sequenced. This huge amount of data can be used to generate protein sequence databases in which each entry corresponds to a mutated peptide associated with certain cancer types. In this chapter, we describe a bioinformatics workflow for creating these databases and detecting mutated peptides in cancer samples from proteomic shotgun experiments. The performance of the proposed method has been evaluated using publicly available datasets from four cancer cell lines.

Keywords

Proteogenomics • TCGA project • SAP detection • Cancer research

A. Garin-Muga
Proteomics and Bioinformatics Unit, Center for Applied Medical Research, University of Navarra, Pamplona, Spain

F.J. Corrales
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

IdiSNA, Navarra Institute for Health Research, Pamplona, Spain

V. Segura (✉)
Proteomics and Bioinformatics Unit, Center for Applied Medical Research, University of Navarra, Pamplona, Spain

IdiSNA, Navarra Institute for Health Research, Pamplona, Spain
e-mail: vsegura@unav.es

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_7

7.1 Introduction

Cancer is one of the leading causes of death worldwide, accounting for 15 % of the total number of annual deaths. Furthermore, in the next two decades, cancer mortality rates are expected to double. There are more than 200 cancer types and each can be classified into several subtypes with different molecular and clinical characteristics (Tomczak et al. 2015). This complexity explains the heterogeneous response to therapy and expected survival rate of patients (McDermott et al. 2011). DNA sequence mutations drive the neoplastic transformation and cause, among other effects, the uncontrolled cell growth in these patients (Hanahan and Weinberg 2000). Therefore, the identification of the complete catalogue of DNA aberrations becomes a priority since it will be the basis for improving not only cancer prevention and its early detection but also its treatment.

The sequencing of the human genome (Lander et al. 2001; Venter et al. 2001) was the first step towards understanding of the complexity of human biology. Over recent years, the development of high-throughput technologies for the characterization of normal and cancer samples has produced an overarching perspective of the human genome and the human proteome (Chin et al. 2011a). For example, gene expression profiling using DNA microarrays (Cordero et al. 2007) has enabled the measurement of the expression level of thousands of genes in a single experiment, resulting in genomic signatures that can be used to classify cancers (Sotiriou and Pusztai 2009; Desmedt et al. 2009). However, the development and deployment of next generation sequencing (NGS) technology have accelerated the discovery of genomic mutations, such as substitutions, insertions, deletions or amplifications (Meyerson et al. 2010; Trapnell et al. 2013). Whole-genome and whole-exome sequencing have proved to be the most valuable methods to identify relevant mutations for the diagnosis and treatment of human disease, including cancer (Pabinger et al. 2014). The identification of germline mutations (Kurian et al. 2014) and the identification of somatic mutations in cancer (Tamborero et al. 2013) are the most common applications for these experiments. The progress in the analysis of cancer genomes has allowed the development of large-scale characterization projects, including the Cancer Genome Project (CGP, http://www.sanger.ac.uk/genetics/CGP), the International Cancer Genome Consortium (ICGC, https://dcc.icgc.org) and The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov). Although the datasets generated in these projects are publically available, access to certain specific information is controlled for the protection of patient privacy (Chin et al. 2011b). Since it started in 2006, the TCGA project has characterized more than 10,000 tumor samples from 34 different cancer types, using high-throughput sequencing. The bioinformatics data analyses of this huge amount of data have resulted in the identification of more than ten million mutations.

Proteogenomics is an emerging field that integrates proteomics, genomics and transcriptomics with the aim of better understanding cellular functions (Faulkner et al. 2015; Nagaraj et al. 2011). Peptides identified in MS-based experiments are aligned against genomic sequence datasets to verify existing gene model annotations or identify novel genes (Ansong et al. 2008). This method has also been applied to the study of the proteome of non-model species, where the availability of a reference protein database is limited (Evans et al. 2012). However, the major advances in proteogenomics have been made in cancer research, and more specifically in the detection of tumor-specific changes in the proteome. Changes such as single amino acid polymorphisms (SAPs) may result in disease initiation, progression or variation in response to treatment (Alfaro et al. 2014; Zhang et al. 2014). The critical stage in the analysis of cancer samples is the generation of the customized peptide databases required to perform the MS searches. Different approaches based on the processing of genomic data, commonly using DNA-Seq or RNA-Seq experiments, have been described (Woo et al. 2014; Nesvizhskii 2014; Wang and Zhang 2013). In this context, the large number of experiments available in projects such as the TCGA is a particularly powerful resource to consolidate this methodology in the clinical oncology setting. The Human Proteome Project (HPP) initiative (Legrain et al. 2011) has played a pioneering role in the integration of transcriptomics and proteomics data (Paik and Hancock 2012; Segura et al. 2014) and the development of new proteogenomics methods and bioinformatics tools is currently an area of active work in the consortium (Krasnov et al. 2015; Nagaraj et al. 2015; Tabas-Madrid et al 2015).

7.2 Bioinformatics Resources for SAP Detection in Cancer

In this section, we propose a bioinformatics workflow to generate proteomic databases for the purpose of identifying SAPs in cancer samples based on the information on somatic mutations obtained from the TCGA project (Fig. 7.1).

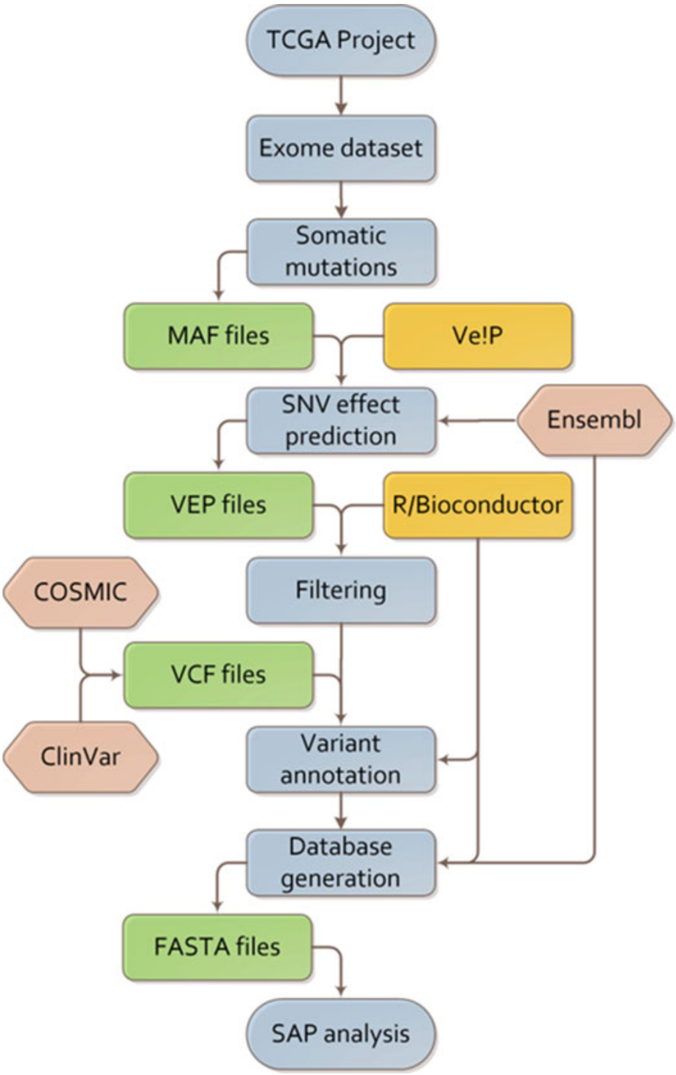


Fig. 7.1 Overall scheme of the SAP detection pipeline using the TCGA exome data. Tasks of the analysis (*blue*), file formats (*green*), databases (*red*) and bioinformatic tools (*orange*) used are shown

7.2.1 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a collaborative project between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The main objective of the project is to describe the key genomic changes in different types of cancers in order to understand the molecular basis of the disease (Tomczak et al. 2015). To achieve this goal, high-throughput technologies have been applied, including microarrays and next generation sequencing (NGS). Since its launch in 2006, the TCGA has analyzed matched tumor and normal tissues from 11,000 patients to study 34 cancer types and subtypes. The availability of these cancer genomic datasets is expected to improve the diagnosis, treatment and prevention of cancer. The structure of TCGA includes several centers responsible for sample collection and processing, high-throughput sequencing and bioinformatics analysis of the obtained datasets.

As shown in Table 7.1, we included 10,183 exomes from 31 cancer types for the generation of the SAP database. We removed three cancer type samples: MESO (there are no exome data available), LAML (the results are based on the hg18 genome) and STAD (the results are not provided using the standard MAF file). The somatic mutation data (MAF files) were downloaded from the data portal of the TCGA project and were processed using in-house R scripts. 3,009,480 unique variants were identified in the dataset without any filtering process. LIHC was the cancer type with the highest number of detected variants (956,761 mutations) and UVM was the cancer type with the fewest number of detected variants (3,918 mutations). The mean number of variants per cancer type was 97,080 mutations and the mean number of variants per analyzed exome was 296.

The SNV calls collected from the TCGA were processed and incorporated into a database of cancer mutations. In order to ensure annotation consistency across all exome results, the detected variants were re-annotated using VeP (McLaren et al. 2010) to predict the effect of the mutations. This software converts SNVs to the corresponding amino acid coordinates and retrieves the SIFT (Kumar et al. 2009) and PolyPhen-2 (Adzhubei et al. 2010) scores for each SNV. We used this information to filter the synonymous SNVs using the R/Bioconductor statistical environment. We only retained those SNVs classified as "missense".

After the filtering process, we obtained 1,161,751 unique genomic variants distributed across the cancer types as shown in Fig. 7.2a. SKCM, LUAD and UCEC were the cancer types with the highest number of identified mutations, while CHOL, PCPG and UVM were the cancer types with the fewest. The relationship between the number of genes and the number of mutations detected in each gene is represented in Fig. 7.2b. It is of note that more than 12,000 genes had at least 50 variants. We also analyzed the number of common variants between different tumor types in order to establish a classification of cancers in terms of their mutations. The clustering of the matrix of the number of mutations shared between each pair of cancer types (Fig. 7.2c) clearly shows three groups of tumors and one cancer type (OV) as an outlier. A more detailed analysis of the number of mutated genes per cancer type (Fig. 7.2d) allowed us to identify tumor-specific mutation genes (256 genes) and ubiquitous mutant genes, defined as those genes that are mutated in all the cancer types studied (37 genes).

Finally, we completed the description of the catalogue of somatic mutations considered for our proteogenomic analysis by determining the percentage of samples in which each gene is altered (Fig. 7.3a). Interestingly, we found that most of the genes are mutated in a very low fraction of the dataset. Only 43 genes were mutated in more than 5 % of the TCGA cancer samples (Fig. 7.3b). This list included genes such as: TTN (titin), TP53 (tumor protein p53), PIK3CA (phosphoinositide-3-kinase, catalytic, alpha polypeptide), BRAF (v-raf murine sarcoma viral oncogene homolog B1) and KRAS (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog); all of which have been described as frequently mutated genes in previous analyses of the TCGA dataset (Ciriello et al. 2013; Kandoth et al. 2013).

Table 7.1 Cancer types studied in the TCGA project. For each tumor we include: its acronym, the number of individuals from whom samples were collected, the number of exomes sequenced and the number of mutations detected in the bioinformatics analysis provided by the data portal of the TCGA

Cancer type	Acronym	#cases/#exomes	#mutations
Acute Myeloid Leukemia	LAML	200/150	–
Adrenocortical carcinoma	ACC	80/80	18,052
Bladder Urothelial Carcinoma	BLCA	412/412	158,417
Brain Lower Grade Glioma	LGG	516/516	59,419
Breast invasive carcinoma	BRCA	1,098/1,081	94,008
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	308/305	82,680
Cholangiocarcinoma	CHOL	36/36	7,679
Colon adenocarcinoma	COAD	461/458	117,118
Esophageal carcinoma	ESCA	185/184	81,248
FFPE Pilot Phase II	FPPP	38/38	40,110
Glioblastoma multiforme	GBM	528/512	60,392
Head and Neck squamous cell carcinoma	HNSC	528/510	137,454
Kidney Chromophobe	KICH	66/66	8,391
Kidney renal clear cell carcinoma	KIRC	536/520	42,196
Kidney renal papillary cell carcinoma	KIRP	291/288	30,320
Liver hepatocellular carcinoma	LIHC	377/375	956,761
Lung adenocarcinoma	LUAD	521/517	219,852
Lung squamous cell carcinoma	LUSC	504/497	65,065
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	48/48	16,462
Mesothelioma	MESO	87/0	0
Ovarian serous cystadenocarcinoma	OV	586/536	12,681
Pancreatic adenocarcinoma	PAAD	185/184	56,871
Pheochromocytoma and Paraganglioma	PCPG	179/179	6,818
Prostate adenocarcinoma	PRAD	498/498	24,399
Rectum adenocarcinoma	READ	171/168	32,815
Sarcoma	SARC	261/255	77,698
Skin Cutaneous Melanoma	SKCM	470/470	298,752
Stomach adenocarcinoma	STAD	443/441	-
Testicular Germ Cell Tumors	TGCT	150/150	25,077
Thymoma	THYM	124/123	37,640
Thyroid carcinoma	THCA	507/496	20,684
Uterine Carcinosarcoma	UCS	57/57	13,048
Uterine Corpus Endometrial Carcinoma	UCEC	548/544	203,455
Uveal Melanoma	UVM	80/80	3,918

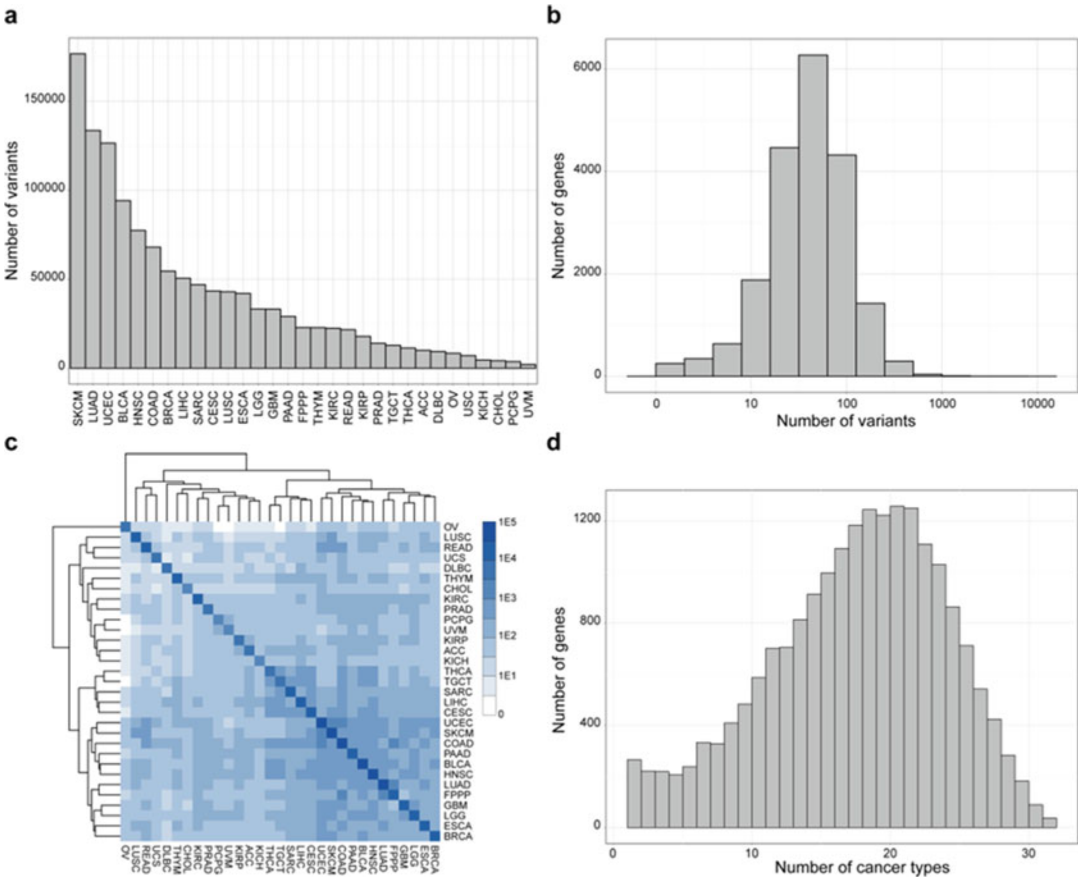


Fig. 7.2 Analysis of the characteristics of the non-synonymous variants detected in the TCGA dataset annotated using VeP. (**a**) Number of variants for each cancer type. (**b**) Number of genes as a function of the number of variants detected in each gene. (**c**) Clustering and heatmap of the matrix of common variants between cancer types. (**d**) Number of mutated genes as a function of the number of cancer types in which a gene is mutated

The functional analysis of this gene set using Ingenuity software (www.ingenuity.com) allowed us to identify enriched functional categories such as Cancer, Cell Cycle, Cell Death and Survival, DNA Repair or Cellular Response to Therapeutics, which confirmed the implication of these genes in cancer-related pathways.

7.2.2 SAP Database Generation

The mutated peptides of the SAP database were generated from the non-synonymous variants of the TCGA samples using the following strategy (Yang and Lazar 2014). First, sequences of 80 amino acids around the mutated amino acid were extracted from the FASTA file of protein sequences (Ensembl version 75). We verified the original amino acid and the position of the mutation, filtering out those cases where a discrepancy was detected. Only the mutated peptides whose sequences did not exist in the reference proteome were stored in the database. A similar workflow was successfully used in the implementation of the dasHPPboard (Tabas-Madrid et al. 2015), a bioinformatics tool that provides access to resources for proteogenomics analyses based on ENCODE (ENCODE Project Consortium et al. 2011) and Illumina Human Body Map (HBM) datasets.

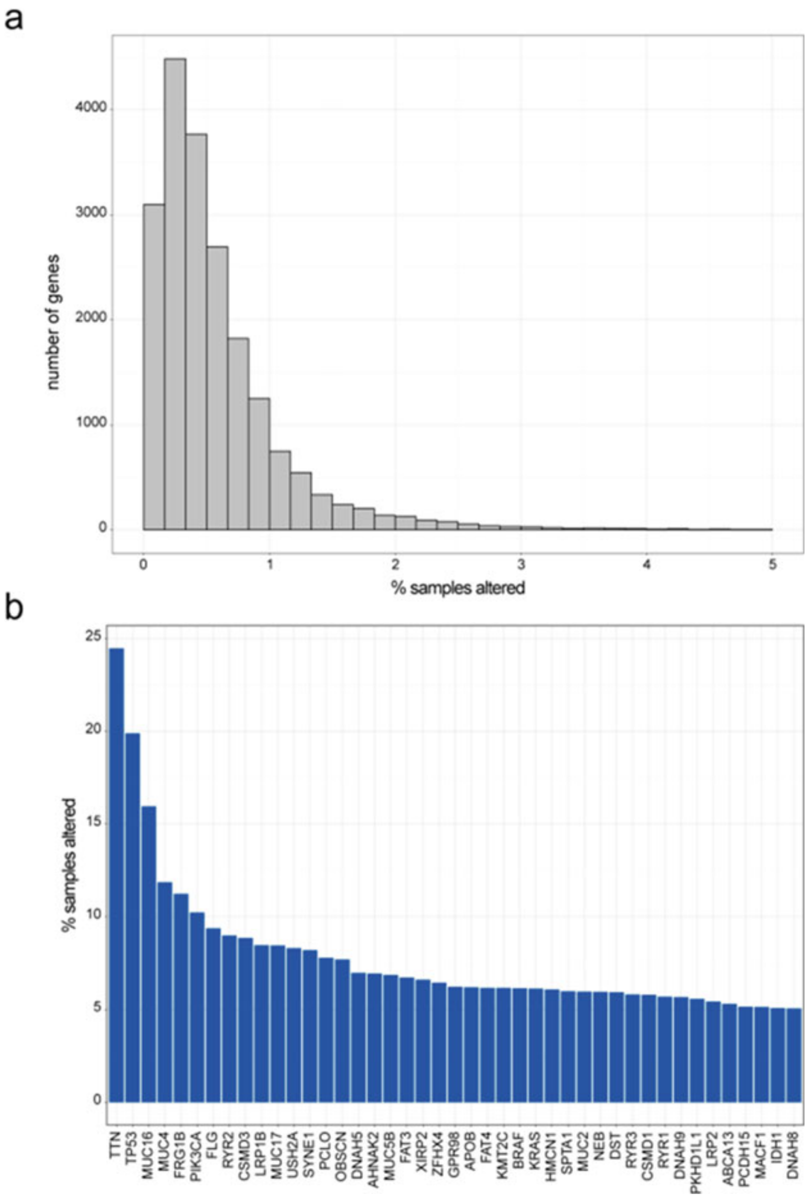


Fig. 7.3 (**a**) Distribution of genes as a function of the percentage of the TCGA samples in which they are mutated. (**b**) Most frequently mutated genes found include well-known oncogenes and tumor suppressors

We identified a total of 1,525,055 unique SAPs located in 92,412 proteins (corresponding to 19,925 genes). The mean number of SAPs per gene was 77, with SKCM being the cancer type with the maximum number of SAPs (222,665 mutated peptides) and UVM the cancer type with minimum number of SAPs (2,738 mutated peptides). The mutation redundancy from the database was removed before the creation of the FASTA file needed to perform the protein searches. Numerical identifiers were used to name the non-redundant peptides of this database, offering the possibility of customizing the content of the FASTA file generated. After searching, the results were easily annotated with in house programmed R scripts.

7.2.3 Annotation of Genetic Variants

Once we completed the database of mutations, a bibliographic and clinical annotation of the genetic variants was incorporated into the analysis workflow. This information can be used to complement the output of VeP and prioritize the detected variants. Two databases of reliable genetic data were selected and processed to facilitate the interpretation of the results: COSMIC (Forbes et al. 2015) and ClinVar (Landrum et al. 2014). COSMIC (Catalog Of Somatic Mutations In Cancer, http://cancer.sanger.ac.uk) is the most complete resource of somatic mutations in human cancer. The release included in the analysis pipeline (v74; September 2015) described 3,480,051 coding mutations in over one million tumor samples and across most human genes. ClinVar (http://www.ncbi.nlm.nih.gov/clinvar) is a database of reports of relationships among genetic variants and phenotypes closely connected with dbSNP (Smigielski et al. 2000) and dbVar (Lappalainen et al. 2013). The incorporated release (November 2015) contained 134,321 variations in 26,406 genes.

We were able to annotate 57.59 % of the non-synonymous variants of the TCGA, 57.47 % using COSMIC and 0.39 % using ClinVar (Fig. 7.4a). Focusing on the ClinVar annotations, we found that 25.36 % of the variants previously described in this database were pathogenic or likely pathogenic, while 28.76 % were benign or likely benign (Fig. 7.4b).

7.3 Proteogenomics Methods for the Identification of SAPs Using Shotgun Experiments

The analysis of high throughput proteomic experiments to detect mutated peptides in cancer samples was performed on the basis of the FASTA databases created from the TCGA datasets. This section is devoted to the proteogenomics study of four cancer cell lines using public shotgun experiments. We used different combinations of search engines and SAP databases to verify that the proteogenomics approach described in this chapter is feasible and is able to provide consistent results.

7.3.1 Public Shotgun Experiments

We analyzed four human cell line experiments available in the PRIDE database (Vizcaíno et al. 2013). These experiments were submitted by the Spanish HPP consortium to the ProteomeXchange repository (http://www.proteomexchange.org/) with accession numbers PXD000039, PXD000442, PXD000443 and PXD000449. We selected 2 replicates for each cell line: Jurkat (human T cell lymphoblast-like cell line), CCD18 (human colon fibroblast cell line), MCF7 (human breast adenocarcinoma cell line) and Huh7 (human hepatocellular carcinoma cell line). Mascot generic files (mgf) were downloaded from the database for further analyses.

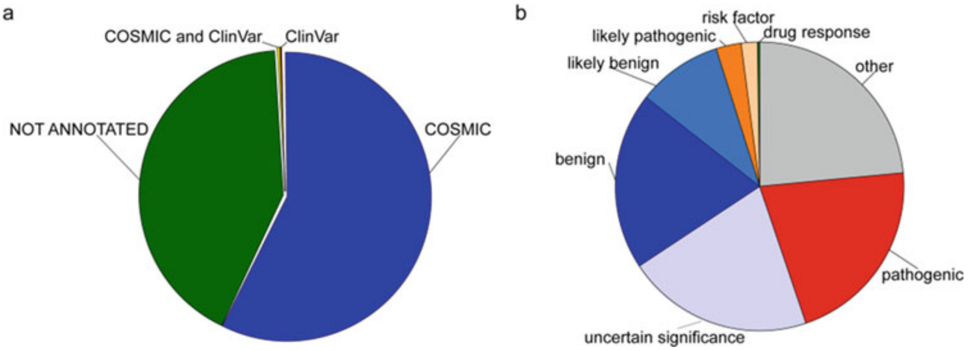


Fig. 7.4 (**a**) Annotation of the mutated peptides incorporated to the SAP database. (**b**) Classification of the mutated peptides annotated using ClinVar

7.3.2 Shotgun Data Analysis

For each sample, two different proteomic analyses were performed: one using the protein reference database (ProteinDB), which in our case is the Ensembl protein database, and the other using the previously generated database of mutated peptides (SAPDB) as shown in Fig. 7.5. In the first search, we identified the proteins present in the samples by applying a FDR criterion at the PSM and protein level. After removal of the assigned spectra from the mgf files, we performed a new search to detect mutated peptides using a SAP database. In this stage, the statistical analyses of the results were carried out at PSM level only.

7.3.2.1 Protein Identification

We searched all the mgf files downloaded from PRIDE against the ProteinDB database (Ensembl version 75) using the target-decoy strategy with an in-house Mascot Server v. 2.3 (Matrix Science, London, U.K.) and Comet (Eng et al. 2013) search engines. A decoy database was created using the peptide pseudo-reversed method and separate searches were performed for target and decoy databases. Search parameters were set as follows: carbamidomethyl cysteine as a fixed modification and oxidized methionine as variable modification. Precursor and fragment mass tolerance were set to 10 ppm and 0.05 Da respectively, and 2 missed cleavages were allowed. False Discovery Rates at the PSM and protein levels using Mayu (Reiter et al. 2009) were calculated. Protein identifications were obtained applying the criteria of PSM FDR < 1 % and protein FDR < 1 % following the C-HPP guidelines.

Protein inference for Mascot results was performed using the PAnalyzer algorithm (Prieto et al. 2012), and non-conclusive protein groups were discarded. In the case of Comet, we considered the protein accessions provided by the search engine (Comet only provides one of the entries in the FASTA file in which the peptide was detected). In Table 7.2, we summarize all the results obtained in the analysis of the proteome of the cell lines under study. The percentage of assigned spectra was very low in all the cell lines (below 15 % of the acquired spectra). The number of proteins and genes detected with Comet was significantly lower than the number detected with Mascot, probably due to the different protein inference algorithm used in both cases. Although the results may not be sufficiently comparable, most of the proteins identified in our study were detected using both search engines.

In order to simplify the evaluation and the comparison of the results we transformed the

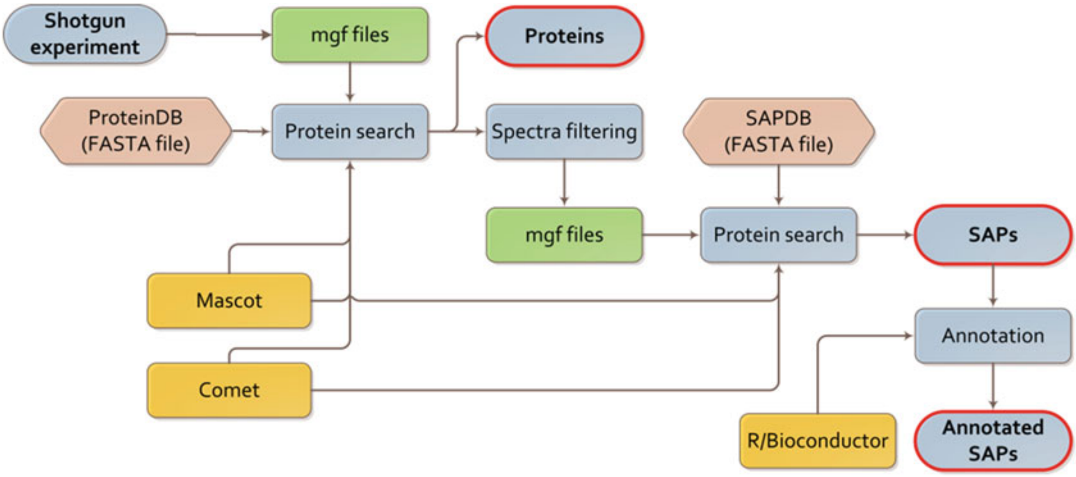
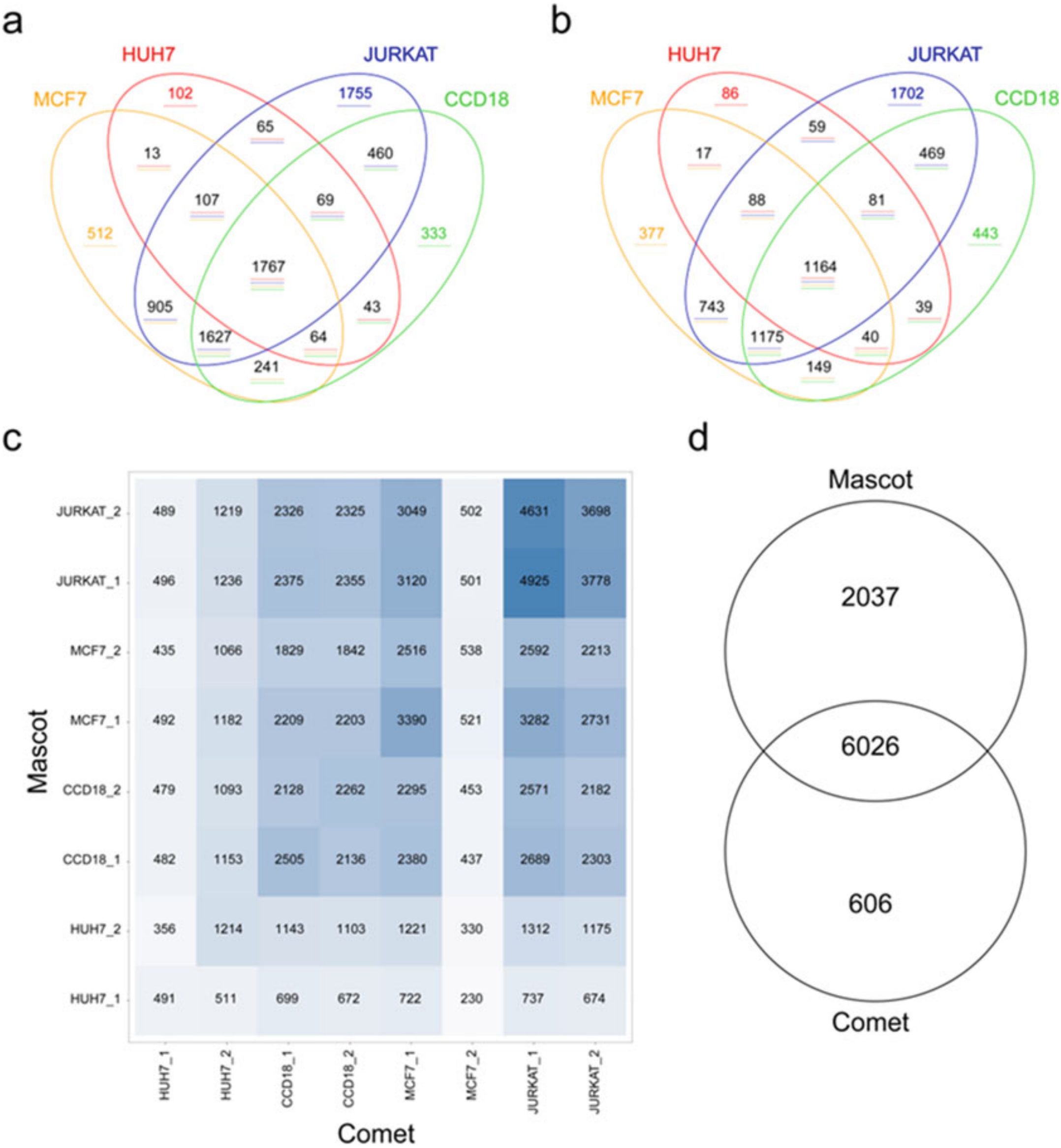


Fig. 7.5 Bioinformatics workflow for SAP detection using proteomic shotgun experiments. Tasks of the analysis (*blue*), file formats (*green*), databases (*red*) and bioinformatic tools (*orange*) used are shown. The *red border* marks the outputs of the workflow

Table 7.2 Number of peptides, proteins and genes identified using ProteinDB (Ensembl version 75) with Mascot and Comet search engines (PSM FDR < 1 %, protein FDR < 1 %)

	HUH7_1	HUH7_2	MCF7_1	MCF7_2	JURKAT_1	JURKAT_2	CCD18_1	CCD18_2
Spectra	589,694	779,849	452,970	468,831	727,488	727,488	831,689	1,492,124
Mascot								
PSMs	27,525	52,389	461,492	130,989	527,701	435,383	538,888	543,506
Peptides	1,367	3,303	28,573	12,175	39,433	36,026	17,981	17,152
Proteins	3,596	6,159	14,000	11,324	19,094	17,888	11,581	11,005
Genes	1,017	1,890	4,785	3,669	6,409	6,110	3,837	3,771
Assigned spectra	4,501	8,444	75,405	22,508	94,976	77,885	80,078	83,025
Comet								
PSMs	2,531	8,332	66,767	2,365	99,214	43,981	72,087	38,349
Peptides	759	3,176	25,418	1,227	39,142	20,922	15,836	9,842
Proteins	541	1,473	4,212	560	6,218	4,413	2,951	3,170
Genes	524	1,394	3,722	548	5,357	3,990	2,680	2,897
Assigned spectra	2,518	7,482	65,550	2,349	97,487	43,401	69,987	37,701

protein accession codes from Ensembl (ENSP) into gene accessions (ENSG). The analysis with Mascot detected 8,063 different genes, 21.92 % common to the 4 cell lines and 33.51 % specific to only one (Fig. 7.6a). On the other hand, the analysis with Comet detected 6,632 genes, 17.55 % common genes and 39.32 % specific genes (Fig. 7.6b). As expected, the comparison between the sets of genes detected in each sample with Mascot and Comet showed that the number of shared identifiers is greater for those samples with the same biological source (Fig. 7.6c).



Mascot \ Comet	HUH7_1	HUH7_2	CCD18_1	CCD18_2	MCF7_1	MCF7_2	JURKAT_1	JURKAT_2
JURKAT_2	489	1219	2326	2325	3049	502	4631	3698
JURKAT_1	496	1236	2375	2355	3120	501	4925	3778
MCF7_2	435	1066	1829	1842	2516	538	2592	2213
MCF7_1	492	1182	2209	2203	3390	521	3282	2731
CCD18_2	479	1093	2128	2262	2295	453	2571	2182
CCD18_1	482	1153	2505	2136	2380	437	2689	2303
HUH7_2	356	1214	1143	1103	1221	330	1312	1175
HUH7_1	491	511	699	672	722	230	737	674

Fig. 7.6 Summary of the ProteinDB search results of the shotgun experiments in the MCF7, CCD18, Jurkat and Huh7 cell lines at protein level. (**a**) Genes identified in the queries performed with Mascot search engine against the database ProteinDB. (**b**) Genes identified in the queries performed with Comet search engine against the database ProteinDB. (**c**) Graphical representation of the genes identified by both search engines and across different cell lines. (**d**) Shared and non-shared genes between the set of proteins obtained in Mascot and Comet analyses

Taking into account all the protein identifications, the percentage of genes detected with both search engines was as high as 69.51 % even though the protein groups were calculated using different methods (Fig. 7.6d).

7.3.2.2 SAP Detection

The detection of mutated peptides started with the removal of the spectra assigned to proteins from the original mgf files. The new datasets were analyzed using the complete SAP database (TCGA) which contained 1,525,055 unique SAPs and cancer-specific databases. Thus, four additional databases were created: a LICH database with 67,694 unique SAPs for the analysis of Huh7, a COAD database with 89,771 unique SAPs for the analysis of CCD18 and a BRCA database with 72,249 unique SAPs for the analysis of MCF7. In the case of the Jurkat cell line only the complete TCGA database was used due to the lack of a specific cancer dataset.

Filtered mgf files were searched using Mascot and Comet with the same query parameters used for the identification of proteins. However, we applied only PSM FDR < 1 % as a statistical threshold for detecting mutated peptides. The results obtained in each of the analysis performed are summarized in Table 7.3.

Despite the large number of spectra retained upon filtering, the results achieved in terms of

Table 7.3 Number of mutated peptides, proteins and genes identified using SAPDB (TCGA or tumor-specific database) with Mascot and Comet search engines (PSM FDR < 1 %)

Search engine	SAP Database	Sample	Assigned spectra	PSMs	SAPs	Mutated proteins	Mutated genes
Comet	BRCA	MCF7_1	83	84	66	185	65
		MCF7_2	136	137	122	352	115
	COAD	CCD18_1	44	52	14	63	14
		CCD18_2	467	475	223	571	179
	LIHC	HUH7_1	1	1	1	4	1
		HUH7_2	446	446	260	704	235
	TCGA	CCD18_1	909	941	241	578	200
		CCD18_2	2,649	2,781	1,051	1,474	481
		HUH7_1	37	38	18	44	18
		HUH7_2	553	567	318	770	242
		JURKAT_1	407	426	234	552	193
		JURKAT_2	121	123	82	216	78
		MCF7_1	426	435	183	454	151
		MCF7_2	5	5	5	14	5
Mascot	BRCA	MCF7_1	81	115	97	218	68
		MCF7_2	18	22	8	24	5
	COAD	CCD18_1	35	65	19	64	10
		CCD18_2	52	82	33	78	21
	LIHC	HUH7_1	4	5	5	9	4
		HUH7_2	6	6	5	9	5
	TCGA	CCD18_1	596	907	316	581	179
		CCD18_2	893	1,238	466	841	256
		HUH7_1	90	134	50	107	35
		HUH7_2	117	153	105	225	70
		JURKAT_1	656	1,111	690	1,150	332
		JURKAT_2	398	554	364	795	239
		MCF7_1	565	1,972	1,627	1,093	343
		MCF7_2	16	27	22	63	12

PSM and identified features (mutated peptides, proteins and genes) were very scarce. These results were consistent with the expected low abundance of mutated peptides in the samples and the random nature of MS experiments. For each selection of search engine (Mascot or Comet) and SAP database (TCGA or tumor-specific), we compared the findings (mutated peptides or genes) for each cell line (Fig. 7.7). In nearly all cases, the number of detected features was higher when the search was performed with Comet. Moreover, it is important to highlight the higher number of identifications obtained using the complete SAPs catalogue than when using cancer-specific databases. What is even more remarkable is the decreased number of mutated genes in relation to the corresponding number of SAPs, although this is likely due to the number of genetic variants per gene.

A question that also deserved our attention is the degree of similarity between the search results obtained using the complete TCGA database and the tumor-specific databases (Fig. 7.8). As expected, for a given cell line most of the mutated peptides were identified using both databases. However, some of the SAPs are shared with another cell line but they were not detected using both databases. This may be due to the effect that the size of the database has on the FDR calculation and the presence of the same genetic variants in different cancer types.

Similarly, we compared the results between search engines taking all the identifications as a whole (Fig. 7.9a–d). The percentage of common SAPs using the TCGA database was 11.7 %, while the percentage of common SAPs using tumor-specific databases was 3.48 %. If the comparison was carried out using the mutated genes detected, the overlap increased both when the TCGA database was used (32.59 %) and also when the tumor-specific databases were used (6.25 %). However, the number of matches in tumor-specific queries was lower than the number of matches in the complete TCGA database. This result suggests that the quality of the peptide identifications obtained with the complete TCGA database was better. In order to confirm this hypothesis, we represented the distributions of the search engine scores (ion score for Mascot and XCorr for Comet). As shown in Fig. 7.9e and 7.9f, the distributions of scores for the searches performed with the TCGA database are slightly higher than the distributions of scores for tumor-specific queries.

In summary, we identified 4,826 mutated peptides in the four cell lines using duplicated shotgun experiments, 2,916 with Mascot and 2,419 with Comet. This can be translated into the detection of 1,580 mutated genes, 877 genes detected with Mascot and 1,161 genes detected with Comet. We were able to annotate 2,422 identified genetic variants using COSMIC (58.04 % of all detections), while only 30 detected genetic variants (0.69 %) were previously described in ClinVar.

7.3.2.3 Functional Analysis of Mutated Genes

The list of mutated genes obtained after the detection of SAPs was further analyzed in order to verify their implication in cancer. Enrichment analysis of disease categories for the 1,580 mutated genes detected was performed using Ingenuity. Of this gene set, 1,252 were associated with the cancer category with a p-value < 1e-30. Furthermore, among the enriched diseases with at least 300 annotated genes we found “breast or colorectal cancer” (663 genes), “colorectal cancer” (555 genes), “liver tumor” (523 genes), “hepatocellular carcinoma” (502 genes), and “hematological neoplasia” (353 genes) or very closely related categories (Fig. 7.10a). This is an important result, which shows that the identified variants are indeed related to the cancer types studied through the public shotgun experiments of cell lines (BRCA, COAD, LALM and LIHC).

We selected the mutated genes detected in at least 7 of the different analyses performed to generate the heatmap and clustering represented in Fig. 7.10b. As expected, the number of genes was higher when the complete TCGA database was used and the groupings were related to the cell line analyzed.

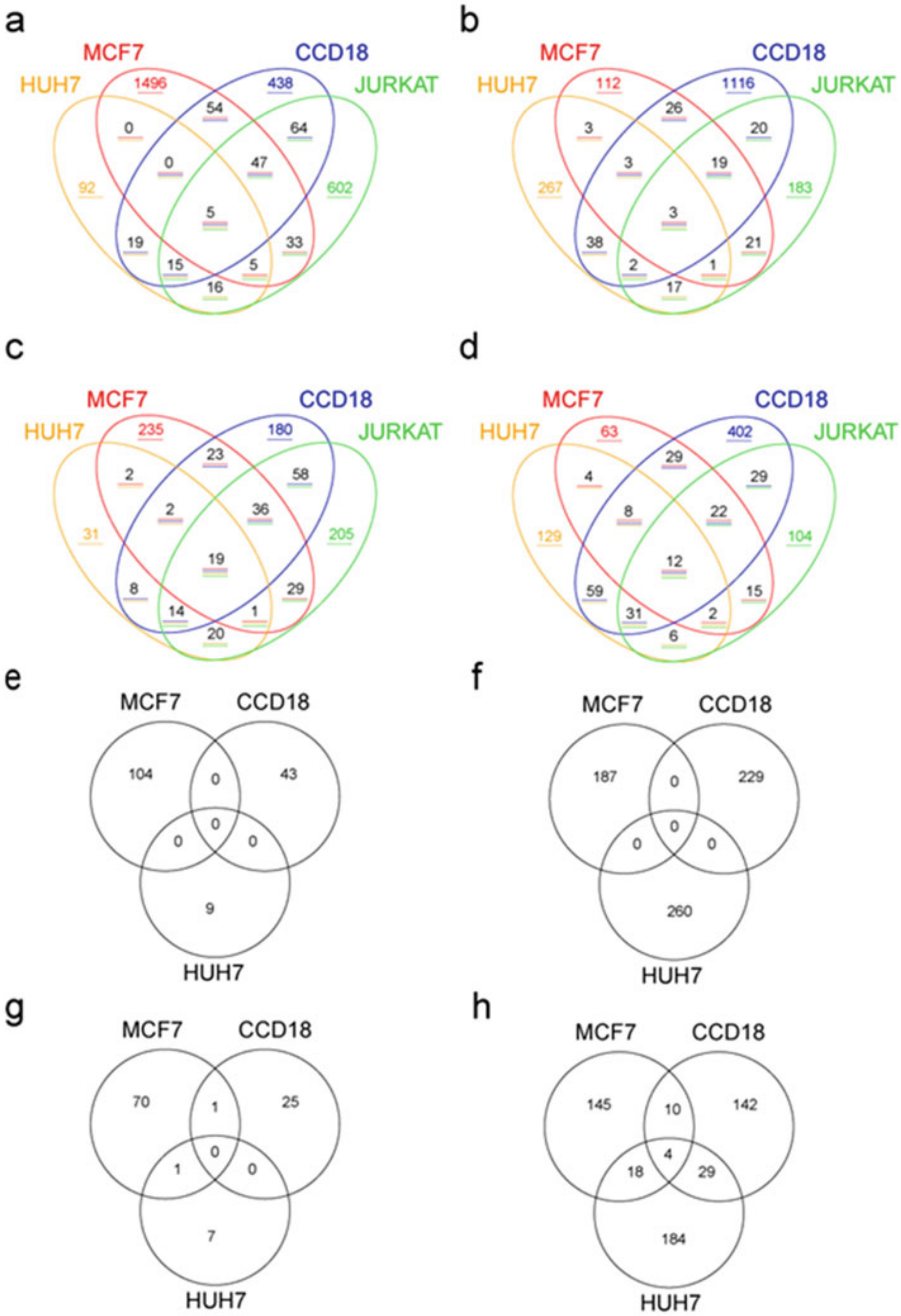


Fig. 7.7 Venn diagrams representing the results of the identification of SAPs and mutated genes using shotgun proteomic data of MCF7, Jurkat, CCD18 and Huh7 cell lines. (**a**) Number of SAPs using the Mascot search engine with TCGA database. (**b**) Number of SAPs using the Comet search engine with TCGA database. (**c**) Number of mutated genes using the Mascot search engine with TCGA database. (**d**) Number of mutated genes using the Comet search engine with TCGA database. (**e**) Number of SAPs using the Mascot search engine with tumor-specific databases. (**f**) Number of SAPs using the Comet search engine with tumor-specific databases. (**g**) Number of mutated genes using the Mascot search engine with tumor-specific databases. (**h**) Number of mutated genes using the Comet search engine with tumor-specific databases

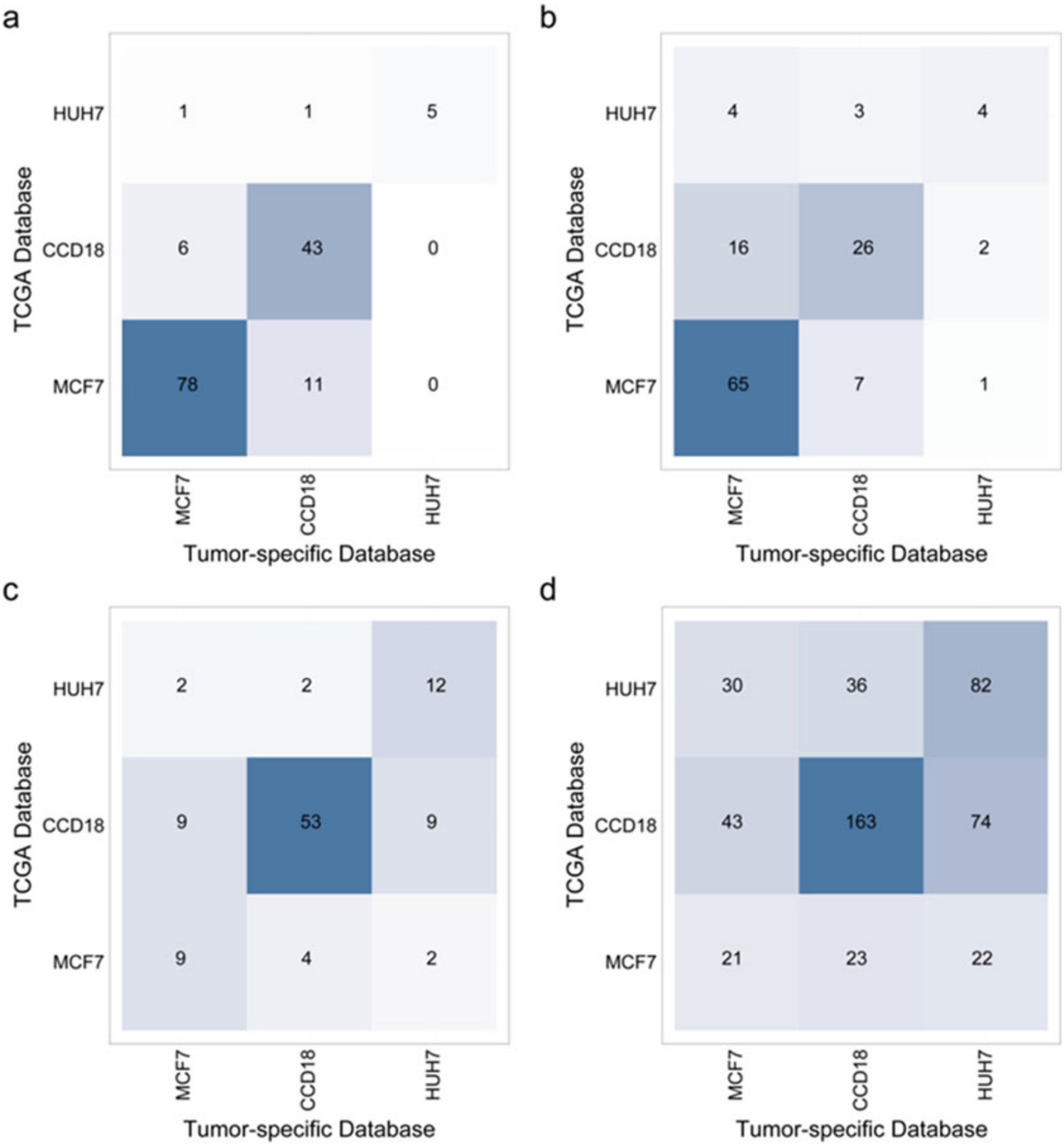


Fig. 7.8 Heatmap representing the similarity among the landscape of mutations obtained using a complete TCGA database or tumor-specific databases. (**a**) Number of SAPs shared among cell lines using the Mascot search engine. (**b**) Number of mutated genes shared among cell lines using the Mascot search engine. (**c**) Number of SAPs shared among cell lines using the Comet search engine. (**d**) Number of mutated genes shared among cell lines using the Comet search engine

7.3.2.4 Curated Analysis of Mutated Peptides

Finally, we completed the analysis of the mutated peptides detected with a manual curation of the best PSM candidates. We selected 35 PSMs obtained using Mascot (the best 5 PSMs for each result shown in Fig. 7.9e). The corresponding spectra were evaluated in a blind manner by two independent MS experts who graded their quality as high, medium or low according to three char-

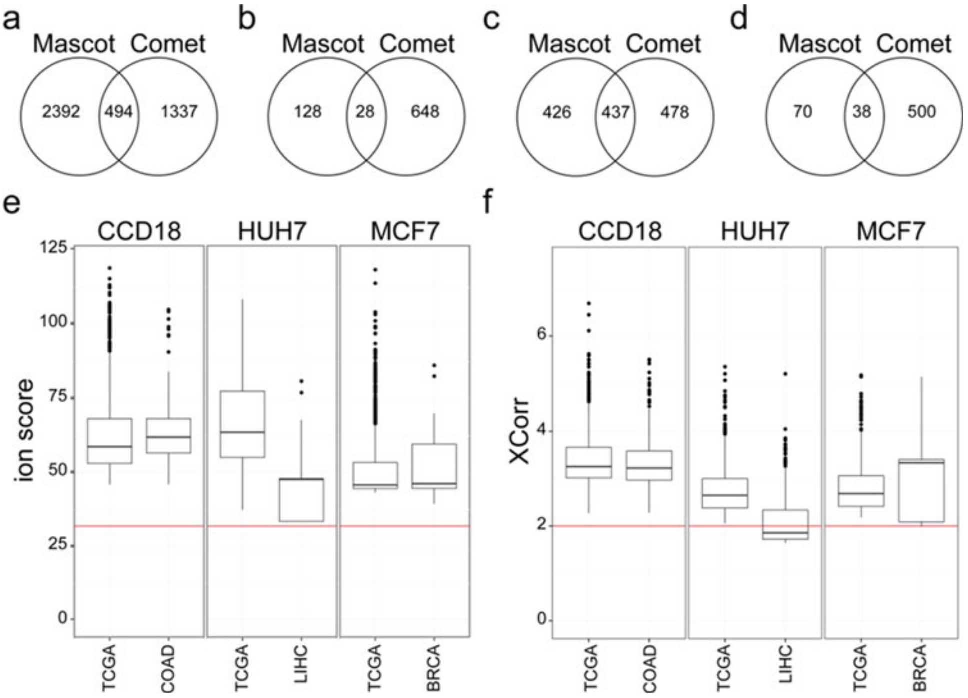


Fig. 7.9 Comparison of the SAP detection analysis as a function of the search engine. (**a**) SAPs identified using the complete TCGA database. (**b**) SAPs identified using tumor-specific databases. (**c**) Mutated genes detected using the complete TCGA database. (**d**) Mutated genes detected using tumor-specific databases. (**e**) Distributions of ion scores for the PSMs obtained in Mascot searches in both TCGA and tumor-specific databases for CCD18, Huh7 and MCF7 cell lines (the *red line* marks the threshold of ion score = 32). (**f**) Distributions of XCorr for the PSMs obtained in Comet searches in both TCGA and tumor-specific databases for CCD18, Huh7 and MCF7 cell lines (the *red line* marks the threshold of XCorr = 2)

acteristics: the quality of y-ion or b-ion series, peak intensities and the presence of no assigned peaks (Jumeau et al. 2015).

The result of this validation process is summarized In Table 7.4. 65.71 % of the spectra were evaluated as “medium” or “high” by the mass spectrometry experts, while the 34.29 % of the spectra were considered of “low” quality. This suggests that a considerable percentage of the detected matches could be considered for further validation steps using other approaches, for example a SRM/MRM assay.

As an example, we chose two of the mutated peptides detected from Table 7.4 (gene APEX1 in COAD cancer type and SERPINB11 gene in LIHC cancer type) to inspect and assign the fragment ions using the spectrum raw signal (Fig. 7.11). In both cases, the y-ion or b-ion series allowed us to establish the sequence of the peptide obtained with the search engine. It is important to emphasize that both sequences contained the mutated amino acid, which validated the presence of the SAPs in CCD18 and Huh7 cell lines respectively.

7.4 Summary

In this chapter, a bioinformatics analysis workflow for the detection of mutated peptides in cancer samples using different computational resources and databases is described. The workflow was divided into two parts. First, the database of SAPs was generated using the information relative to the variants of interest. After the FASTA files required for peptide identifications were generated, the rest of the analysis was aimed at the statistical analysis and the annotation of the results of the proteomic searches.

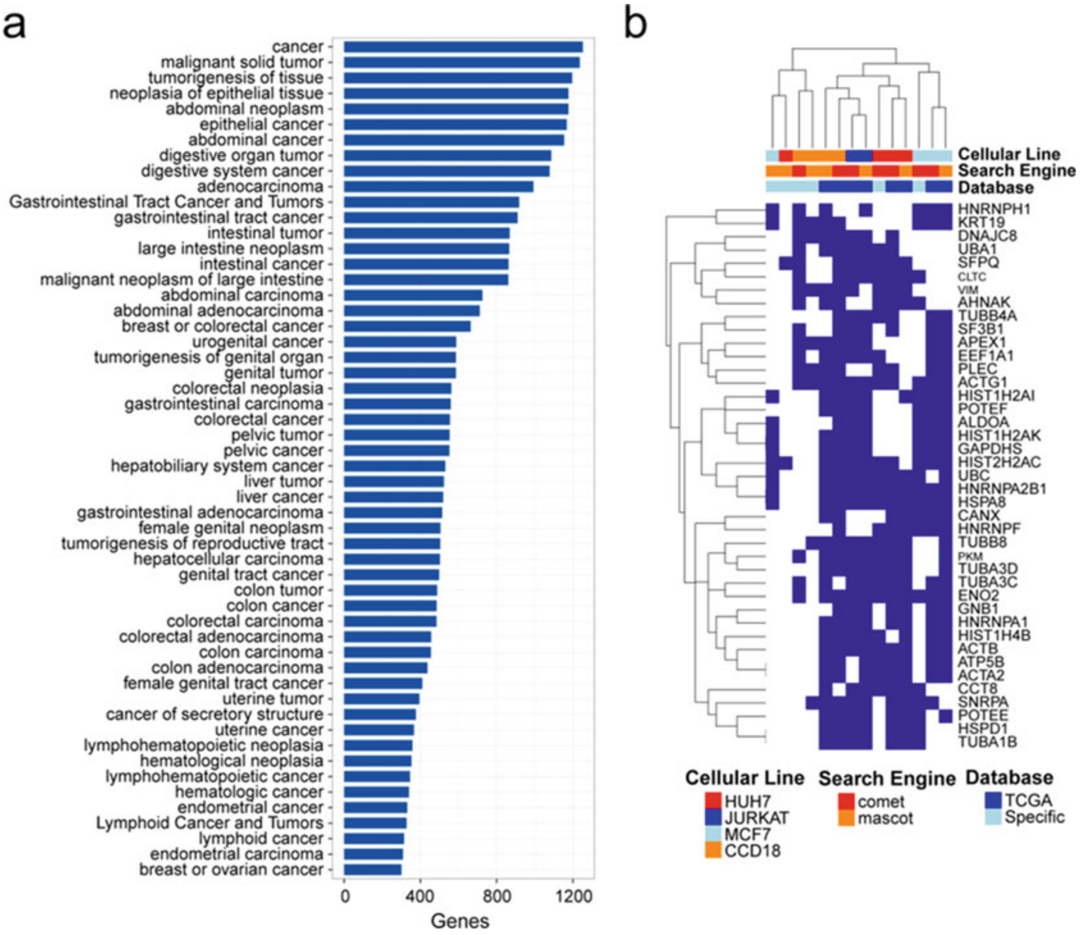


Fig. 7.10 (**a**) Enrichment of disease categories using Ingenuity for the list of 1580 mutated genes detected in the proteogenomics analysis of SAPs. (**b**) Clustering of the mutated genes detected in at least 7 of the analyses performed. Each column corresponds to a combination of a specific cell line, search engine and database. *Blue squares* indicate the detection of the mutated gene in the result of the analysis

The database was created from the data obtained from the TCGA project. More than 10,000 exomes from 31 cancer types were analyzed in this project to obtain a catalogue of more than three million somatic variants. The VeP software was used to infer the effect of the variants in order to retain only missense mutations in the SAP database. In addition to that, the variants were also annotated with bibliographic information and clinical significance using two reliable resources of somatic mutations in cancer, the COSMIC and ClinVar databases. The final result of the data processing was a set of mutations that can be used to generate FASTA files on-demand. The selection of mutations could be based on tumor type, clinical significance, or other criteria of interest.

We studied the mutation landscape of 4 cancer cell lines (Huh7, MCF7, CCD18 and Jurkat) to verify the performance of the proposed approach. Publicly available proteomic shotgun experiments were downloaded from PRIDE database and analyzed using a reference proteome (Ensembl version 75) and two SAP databases: all TCGA mutated peptides and tumor-specific mutated peptides. Furthermore, we compared the efficiency of the detection using two commonly used search engines, Mascot and Comet. In a first

Table 7.4 List of the 35 PSMs selected for manual curation

Gene ID	Gene name	Chr	Position	R/A	Cancer type	Spectrum quality
ENSG00000156508	EEF1A1	6	74,228,666	P/A	HNSC	Low
ENSG00000184260	HIST2H2AC	1	149,858,837	Q/E	LUAD	Low
ENSG00000147140	NONO	X	70,514,267	I/T	LIHC	Medium
ENSG00000178209	PLEC	8	145,007,187	A/V	COAD, CESC	Low
ENSG00000180543	TSPYL5	8	98,289,714	T/S	COAD	Medium
ENSG00000168090	COPS6	7	99,688,715	R/C	ACC	Medium
ENSG00000144381	HSPD1	2	198,363,449	Q/K	LUAD	High
ENSG00000197157	SND1	7	127,724,820	A/S	LUAD, GBM	High
ENSG00000122566	HNRNPA2B1	7	26,237,265	W/G	LIHC	Low
ENSG00000101558	VAPA	18	9,950,451	M/I	SARC	Medium
ENSG00000144381	HSPD1	2	198363449	Q/K	LUAD	Low
ENSG00000077312	SNRPA	19	41265363	D/Y	UCEC	High
ENSG00000171858	RPS21	20	60,962,906	K/R	CHOL	Low
ENSG00000144381	HSPD1	2	198,363,449	Q/K	LUAD	Low
ENSG00000077312	SNRPA	19	41,265,363	D/Y	UCEC	High
ENSG00000142168	SOD1	21	33,040,872	V/A	UCEC	Medium
ENSG00000104833	TUBB4A	19	6,502,196	G/S	COAD	High
ENSG00000104833	TUBB4A	19	6,501,408	G/V	LUAD	High
ENSG00000171314	PGAM1	10	99,190,370	P/L	UCS	High
ENSG00000184640	SEPT9	17	75,488,782	V/G	KIRP	Medium
ENSG00000169045	HNRNPH1	5	179,043,881	G/D	BRCA	Medium
ENSG00000105679	GAPDHS	19	36,034,281	D/Y	BRCA	High
ENSG00000169045	HNRNPH1	5	179,043,881	G/D	BRCA	Low
ENSG00000141837	CACNA1A	19	13,356,070	R/C	BRCA	Medium
ENSG00000122566	HNRNPA2B1	7	26,236,568	E/Q	BRCA	Medium
ENSG00000100823	APEX1	14	20,925,154	D/E	COAD	Medium
ENSG00000088682	COQ9	16	57,486,729	G/S	COAD	Low
ENSG00000181873	IBA57	1	228,362,682	G/S	COAD	Medium
ENSG00000100823	APEX1	14	20,925,154	D/E	COAD	High
ENSG00000178209	PLEC	8	145,001,031	H/R	COAD	Medium
ENSG00000116560	SFPQ	1	35,656,352	K/T	LIHC	Low
ENSG00000116560	SFPQ	1	35,656,352	K/T	LIHC	Low
ENSG00000206072	SERPINB11	18	61,387,312	A/T	LIHC	High
ENSG00000184260	HIST2H2AC	1	149,858,769	R/P	LIHC	Low
ENSG00000162396	PARS2	1	55,223,846	G/D	LIHC	Medium

search for each sample we identified the proteins present in the cell lines, using a target-decoy strategy and a FDR threshold of 1 % at the PSM and protein levels. Protein inference was performed applying the PAnalyzer algorithm in Mascot analyses and the output of the search engine in the Comet analyses. Although the results were not entirely comparable, considerable overlap was found.

Finally, after removal of the previously assigned spectra a second search was performed against the SAP database (complete TCGA data-

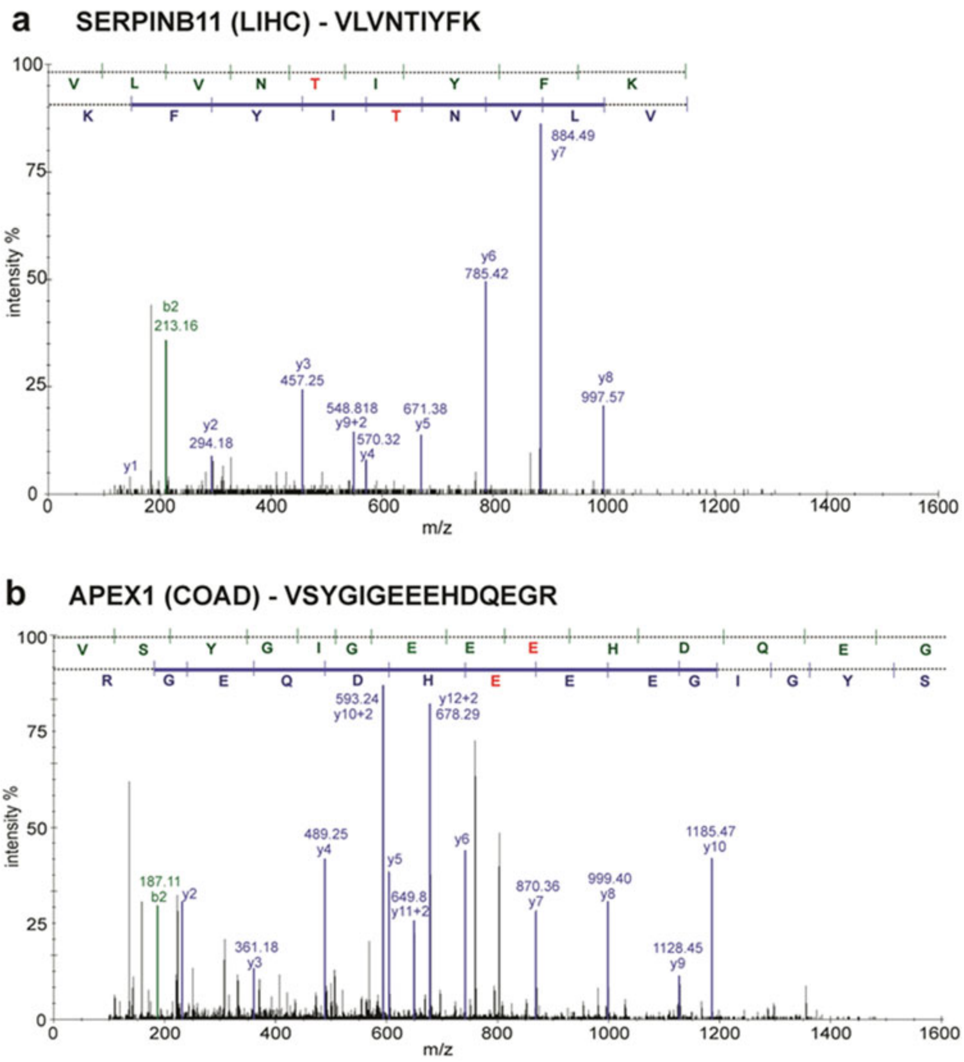


Fig. 7.11 (**a**) Raw spectrum and assigned b-ions and y-ions for the mutated peptide VLVNTIYFK detected in LIHC tumor type in the gene SERPINB11. (**b**) Raw spectrum and assigned b-ions and y-ions for the mutated peptide VSYGIGEEEHDQEGR detected in COAD tumor type in the gene APEX1

set or cancer-specific mutations). The set of mutated genes detected with these analyses were functionally annotated and their implication in cancer, and specifically in the cancer types under study, was verified. Another validation step was performed using two independent experts to evaluate the best PSMs detected. This study confirmed the validity of the proteogenomics approach to detect mutated peptides in cancer samples.

Acknowledgments All participating laboratories are members of ProteoRed-ISCIII. This work was supported by: Carlos III Health Institute of Spain (ISCIII, FIS PI11/02114 and FIS PI14/01538)-Fondos FEDER (EU); grants SAF2014-5478-R from Ministerio de Economía y Competitividad. The CIMA Proteomics Unit belongs to ProteoRed, PRB2-ISCIII, supported by grant PT13/0001 L We also thank the Proteomics, Genomics and Bioinformatics Core Facility of CIMA, especially to Elizabeth Guruceaga, María Mora and Leticia Odriozola for technical support.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248–249.

Alfaro, J. A., Sinha, A., Kislinger, T., et al. (2014). Onco-proteogenomics: Cancer proteomics joins forces with genomics. *Nature Methods, 11*(11), 1107–1113.

Ansong, C., Purvine, S. O., Adkins, J. N., et al. (2008). Proteogenomics: Needs and roles to be filled by proteomics in genome annotation. *Briefings in Functional Genomics & Proteomics, 7*(1), 50–62.

Chin, L., Andersen, J. N., & Futreal, P. A. (2011a). Cancer genomics: From discovery science to personalized medicine. *Nature Medicine, 17*(3), 297–303.

Chin, L., Hahn, W. C., Getz, G., et al. (2011b). Making sense of cancer genomic data. *Genes & Development, 25*(6), 534–555.

Ciriello, G., Miller, M. L., Aksoy, B. A., et al. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics, 45*(10), 1127–1133.

Cordero, F., Botta, M., & Calogero, R. A. (2007). Microarray data analysis and mining approaches. *Briefings in Functional Genomics & Proteomics, 6*(4), 265–281.

Desmedt, C., Sotiriou, C., & Piccart-Gebhart, M. J. (2009). Development and validation of gene expression profile signatures in early-stage breast cancer. *Cancer Investigation, 27*(1), 1–10.

Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *Proteomics, 13*(1), 22–24.

Evans, V. C., Barker, G., Heesom, K. J., et al. (2012). De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature Methods, 9*(12), 1207–1211.

Faulkner, S., Dun, M. D., & Hondermarck, H. (2015). Proteogenomics: Emergence and promise. *Cellular and Molecular Life Sciences, 72*(5), 953–957.

Forbes, S. A., Beare, D., Gunasekaran, P., et al. (2015). COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research, 43*(Database issue), D805–D811.

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

Jumeau, F., Com, E., Lane, L., et al. (2015). Human spermatozoa as a model for detecting missing proteins in the context of the chromosome-centric Human Proteome Project. *Journal of Proteome Research, 14*(9), 3606–3620.

Kandoth, C., McLellan, M. D., Vandin, F., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature, 502*(7471), 333–339.

Krasnov, G. S., Dmitriev, A. A., Kudryavtseva, A. V., et al. (2015). PPLine: An automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics. *Journal of Proteome Research, 14*(9), 3729–3737.

Kumar, P., Henikoff, S., & Pauline, C. N. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols, 4*(8), 1073–1082.

Kurian, A. W., Hare, E. E., Mills, M. A., et al. (2014). Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *Journal of Clinical Oncology, 32*(19), 2001–2009.

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

Landrum, M. J., Lee, J. M., Riley, G. R., et al. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research, 42*(Database issue), D980–D985.

Lappalainen, I., Lopez, J., Skipper, L., et al. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Research, 41*(Database issue), D936–D941.

Legrain, P., Aebersold, R., Archakov, A., et al. (2011). The human proteome project: Current state and future direction. *Molecular and Cellular Proteomics, 10*(7), M111.009993.

McDermott, U., Downing, J. R., & Stratton, M. R. (2011). Genomics and the continuum of cancer care. *New England Journal of Medicine, 364*(4), 340–350.

McLaren, W., Pritchard, B., Rios, D., et al. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics, 26*(16), 2069–2070.

Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics, 11*(10), 685–696.

Nagaraj, N., Wisniewski, J. R., Geiger, T., et al. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology, 7*, 548.

Nagaraj, S. H., Waddell, N., Madugundu, A. K., et al. (2015). PGTools: A software suite for proteogenomic data analysis and visualization. *Journal of Proteome Research, 14*(5), 2255–2266.

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods, 11*(11), 1114–1125.

Pabinger, S., Dander, A., Fischer, M., et al. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics, 15*(2), 256–278.

Paik, Y. K., & Hancock, W. S. (2012). Uniting ENCODE with genome-wide proteomics. *Nature Biotechnology, 30*(11), 1065–1067.

Prieto, G., Aloria, K., Osinalde, N., et al. (2012). PAnalyzer: A software tool for protein inference in shotgun proteomics. *BMC Bioinformatics, 13*, 288.

ENCODE Project Consortium, Bernstein, B. E., Birney, E., et al. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology, 9*(4), e1001046.

Reiter, L., Claassen, M., Schrimpf, S. P., et al. (2009). Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular and Cellular Proteomics, 8*(11), 2405–2417.

Segura, V., Medina-Aunon, J. A., Mora, M. I., et al. (2014). Surfing transcriptomic landscapes. A step beyond the annotation of chromosome 16 proteome. *Journal of Proteome Research, 13*(1), 158–172.

Smigielski, E. M., Sirotkin, K., Ward, M., et al. (2000). dbSNP: A database of single nucleotide polymorphisms. *Nucleic Acids Research, 28*(1), 352–355.

Sotiriou, C., & Pusztai, L. (2009). Gene-expression signatures in breast cancer. *New England Journal of Medicine, 360*(8), 790–800.

Tabas-Madrid, D., Alves-Cruzeiro, J., Segura, V., et al. (2015). Proteogenomics dashboard for the Human Proteome Project. *Journal of Proteome Research, 14*(9), 3738–3749.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Science Reports, 3*, 2650.

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology (Pozn), 19*(1A), A68–A77.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., et al. (2013). Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nature Biotechnology, 31*(1), 46–53.

Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001). The sequence of the human genome. *Science, 291*(5507), 1304–1351.

Vizcaíno, J. A., Côté, R. G., Csordas, A., et al. (2013). The PRoteomics IDEntifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Research, 41*(Database issue), D1063–D1069.

Wang, X., & Zhang, B. (2013). customProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics, 29*(24), 3235–3237.

Woo, S., Cha, S. W., Na, S., et al. (2014). Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics, 14*(23–24), 2719–2730.

Yang, X., & Lazar, I. M. (2014). XMAn: A Homo sapiens mutated-peptide database for the MS analysis of cancerous cell states. *Journal of Proteome Research, 13*(12), 5486–5495.

Zhang, B., Wang, J., Wang, X., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature, 513*(7518), 382–387.

8 Developments for Personalized Medicine of Lung Cancer Subtypes: Mass Spectrometry-Based Clinical Proteogenomic Analysis of Oncogenic Mutations

Toshihide Nishimura and Haruhiko Nakamura

Abstract

Molecular therapies targeting lung cancers with mutated epidermal growth factor receptor (EGFR) by EGFR-tyrosin kinase inhibitors (EGFR-TKIs), gefitinib and erlotinib, changed the treatment system of lung cancer. It was revealed that drug efficacy differs by race (*e.g.*, Caucasians vs. Asians) due to oncogenic driver mutations specific to each race, exemplified by gefitinib / erlotinib. The molecular target drugs for lung cancer with anaplastic lymphoma kinase (ALK) gene translocation (the fusion gene, EML4-ALK) was approved, and those targeting lung cancers addicted ROS1, RET, and HER2 have been under development. Both identification and quantification of gatekeeper mutations need to be performed using lung cancer tissue specimens obtained from patients to improve the treatment for lung cancer patients: (1) identification and quantitation data of targeted mutated proteins, including investigation of mutation heterogeneity within a tissue; (2) exploratory mass spectrometry (MS)-based clinical proteogenomic analysis of mutated proteins; and also importantly (3) analysis of dynamic protein–protein interaction (PPI) networks of proteins significantly related to a subgroup of patients with lung cancer not only with good efficacy but also with acquired resistance. MS-based proteogenomics is a promising approach to directly capture mutated and fusion proteins expressed in a clinical sample. Technological developments are further expected, which will provide a powerful solution for the stratification of patients and drug discovery (Precision Medicine).

T. Nishimura (✉)
Translational Medicine Informatics, St. Mariana University School of Medicine, Kanagawa, Japan

Research & Development, Biosys Technologies Inc., Tokyo, Japan

Center of Excellence in Biological and Medical Mass Spectrometry, BMC, Lund University, Lund, Sweden
e-mail: t-nisimura-tmu@hotmail.co.jp

H. Nakamura
Translational Medicine Informatics, St. Mariana University School of Medicine, Kanagawa, Japan

Chest Surgery, St. Mariana University School of Medicine, Kanagawa, Japan
e-mail: h-nakamura@marianna-u.ac.jp

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_8

Keywords

Lung adenocarcinoma • Clinical proteogenomics • Oncogenic driver mutation • Fusion gene • Mass spectrometry (MS) • Third-generation EGFR tyrosine kinase inhibitor (TKI) • Protein-protein interaction (PPI) network • Drug resistance • Precision medicine

8.1 Introduction

Lung cancer is the most prevalent type of cancer in the world and it is a leading cause of death (Jemal et al. 2011; the data base of Japanese Ministry of Health and Labor and Welfare 2013). In Japan, annual deaths from lung cancer are increasing and currently approach 70,000 (The data base of Japanese Ministry of Health and Labor and Welfare 2013), while in the United States with a recent decreasing trend in mortality, more than 160,000 succumb annually (The data base of National Cancer Institute at the National Institute of Health 2013). Tobacco smoking is the main risk factor for lung cancer; however, approximately 25 % of lung cancers worldwide occur in never-smokers (Ferlay et al. 2010; Siegel et al. 2011). Moreover, the risk of lung cancer differs by race / ethnicity. In the United States, approximately 10 % of patients with lung cancer are never-smokers (Scagliotti et al 2009), while in Asia, >30 % of patients with lung cancer are never-smokers and more than 50 % of lung cancers occur in women who are never-smokers (Toh et al. 2004).

Weinstein (2001) proposed that cancer cells acquire abnormalities in multiple oncogenes and tumor suppressor genes but that the inactivation of a single critical oncogene can induce cancer cells to differentiate into cells with a normal phenotype or to undergo apoptosis (Oncogene Addiction). Only a few driver oncogenic genes would result in their dysfunctions and lead to anti-tumor effects whereas numerous abnormal oncogenes are expressed in a tumor. The successful application of this phenomenon is the molecular therapy targeting lung cancers with mutated epidermal growth factor receptor (EGFR) by EGFR-tyrosine kinase inhibitors (EGFR-TKIs), gefitinib and erlotinib, which thereafter changed the treatment system of lung cancer. Several clinical trials, such as the NEJ002 (Miyauchi et al. 2015) and WJTOG3405 (Mitsudomi et al. 2010) studies, reported the superiority of gefitinib over carboplatin (CBDCA) and paclitaxel (PTX; Taxol) (CBDCA/PTX) on progression-free survival (PFS) as the first-line treatment for advanced non-small cell lung cancer (NSCLC) harboring an epidermal growth factor receptor (EGFR) mutation. In addition, in 2012 the molecular target drug, crizotinib, was approved for lung cancers with an anaplastic lymphoma kinase (ALK) gene translocation (the fusion gene, EML4-ALK) (Camidge and Doebele 2012), and further, those targeting lung cancers addicted ROS1 (Shaw et al. 2012), RET (Sasaki et al. 2012), and HER2 (Mar and Vredenburgh 2015) have also been under development. It seems that such personalized treatments are changing lung cancer to a chronic disease, which might even be able to be cured.

8.2 Lung Cancer Subtypes and Mutations

8.2.1 Lung Cancer Subtypes

Large-cell lung cancer (LCC) is one of the subtypes in which cancerous large cells grow without any distinctive tissue construct. Small-cell lung cancer (SCLC) is the subtype of an aggressive neuroendocrine tumor consisting of small bare nuclei cells. Travis et al. (1991) proposed a new subtype of large-cell lung carcinoma, named large cell neuroendocrine carcinoma (LCNEC) in 1991, and the World Health Organization finally adopted it for the revised pathological classification of lung cancer in the 1999 and 2004 WHO classifications. (Travis et al. 1999, 2004) LCNEC exhibits morphology similar to LCC, but neuroendocrine differentiation like SCLC that could be judged by the expression of at least one of three representative neuroendocrine proteins; CD56, synaptophysin (Syn) or chromogranin A (CGA). The developmental history of the tissue origin is currently unknown for these three types of lung cancers. LCNEC has a poor prognosis, similarly to small-cell lung carcinoma (SCLC), and the survival rate is just 18 % in IA-stage, only by resection (Dresler et al. 1997; Battafarano et al. 2005). Currently, similarly to non-small-cell lung carcinoma (NSCLC), resection is the first choice, followed by adjuvant therapy, as for SCLC. Surgical resection of LCNEC in many series has been described with 5-year actuarial survival that is far worse than that reported for other histological variants of non-small-cell lung cancer (NSCLC). There have been considerable debates on whether these tumors should be classified and treated as NSCLC or SCLC. The large-scale epidemiologic study has compared the presenting and prognostic characteristics of patients with LCNEC with those of patients with SCLC or other large cell carcinomas (OLCs) with respect to overall survival (OS) and lung cancer-specific survival (LCSS) rates for patients undergoing definitive resection without radiotherapy (S-NoRT), and they have concluded that LCNEC should continue to be classified and treated as a large cell carcinoma because the clinical, histopathologic, and biologic features of LCNEC are more similar to OLC than to SCLC (Varlotto et al. 2011; Travis 2010).

8.2.2 Lung Adenocarcinoma Classification

In an increasing trend worldwide, advances in chest high-resolution computed tomography (HRCT) scanning technology have enabled the localization of small adenocarcinoma nodules (Nakamura and Saji 2014) at an earlier and potentially more curable stage of development than previously possible (Koike et al. 2009). There are 90 million current and ex-smokers in the United States who are at increased risk of lung cancer. The published data from the National Lung Screening Trial (NLST) suggest that yearly screening with low dose thoracic CT scans in heavy smokers can reduce lung cancer mortality by 20 % and all-cause mortality by 7 % (The National Lung Screening Trial Research Team 2011).

In 2011, the new pathologic classification of lung adenocarcinoma was proposed by the International Association for the Study of Lung Cancer (IASLC), the American Thoracic Society (ATS) and the European Respiratory Society (ERS) (Travis et al. 2011). In the new classification, the concept of adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA) were newly introduced and the term bronchioloalveolar carcinoma (BAC) was abolished. Additionally, invasive adenocarcinomas were categorized into 6 subtypes, lepidic (LEP), acinar (CAN), papillary (PAP), micropapillary (MP), solid (SLD), and variants, according to the predominant histologic pattern. Both AIS and MIA were defined as tumors ≤ 3 cm in size. AIS is a preinvasive lesion showing pure lepidic growth without invasion. MIA is also lepidic predominant tumor but with ≤ 5 mm invasion. LEP is an invasive adenocarcinoma showing former non-mucinous BAC pattern with > 5 mm invasion. These 3 lepidic type adenocarcinomas are speculated to show stepwise progression from AIS, MIA, to LEP. After complete resection of AIS or MIA, usually 100 % of recurrencefree 5year survival can be obtained (Travis et al. 2011), while some recurrent cases are found after resection of LEP (Yoshizawa et al. 2011, 2013; Gu et al. 2013). Since postoperative prognoses between the AIS plus MIA group and LEP are different, differential protein expressions associated with invasiveness of cancer cells in each subtype should play important roles to determine local recurrences and survivals.

Recently the driver mutations, including the EGFR, KRAS, HER2, BRAF, ALK, RET, and PIK3CA, were investigated across the IASLC/ATS/ERS morphologic classifications in lung adenocarcinoma incorporated with the clinicopathological characteristics to evaluate their mutual correlation, in which lung adenocarcino-

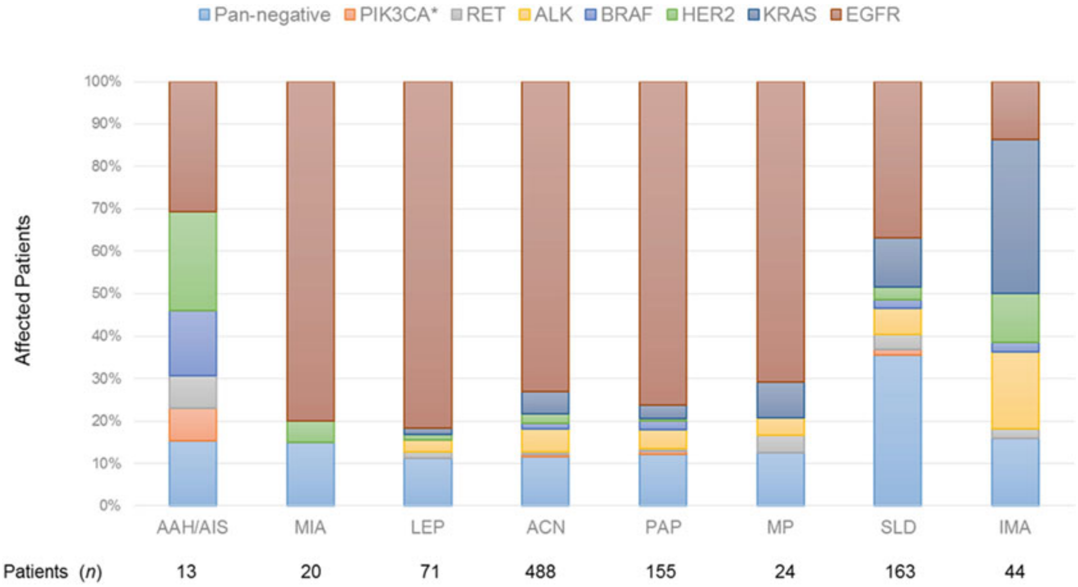


Fig. 8.1 Driver mutation spectrum, according to the novel IASLC/ATS/ERS classification (Hu et al. 2014). Note: *Indicates samples harboring the PIK3CA mutation without overlap with other driver mutations. Abbreviations: *IASLC* International Association for the Study of Lung Cancer, *ATS* American Thoracic Society, *ERS* European Respiratory Society, *AAH* atypical adenomatous hyperplasia, *AIS* adenocarcinoma in situ, *MIA* minimally invasive adenocarcinoma, *LEP* lepidic predominant, *ACN* acinar predominant, *PAP* papillary predominant, *MP* micropapillary predominant, *SLD* solid predominant, *IMA* invasive mucinous adenocarcinoma

mas obtained from 1015 Chinese patients were analyzed (Hu et al. 2014). There, the driver mutations, EGFR (exons 18–22), HER2 (exons 18–21), KRAS (exons 2–3), BRAF (exons 11–15), and PIK3CA (exons 9–20) were amplified by using the polymerase chain reaction (PCR) with cDNA used for Sanger sequencing. The ALK and RET rearrangements were screened by using PCR and quantitative real-time PCR with cDNA and confirmed with fluorescence in-situ hybridization in formalin-fixed paraffin-embedded (FFPE) tissue specimens. Figure 8.1 demonstrated that all driver mutations across the IASLC/ATS/ERS classifications were mutually exclusive except in 18 patients coexisting EGFR and PIK3CA mutations, 4 with both the KRAS and PIK3CA mutations, and 1 with both the RET and PIK3CA mutations. The EGFR mutation was 64.7 % in overall frequency, much higher than in the Caucasian population, and the KRAS mutation was 7.1 %, much lower than in Caucasian patients.

Never-smoker East Asian females have a tendency to develop adenocarcinoma, and these never-smokers exhibit higher treatment response rates to epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKIs), such as gefitinib (Iressa™) and erlotinib (Tarceva™), than those with a history of tobacco smoking (Ha et al. 2015). EGFR-TKIs block EGFR phosphorylation and subsequent signal transduction pathways involved in proliferation, metastasis, angiogenesis and apoptosis inhibition, and gefitinib is the first molecular targeting drug significantly beneficial for Asian lung cancer patients. Figure 8.2 shows the frequency of driver gene mutations in lung adenocarcinomas from East Asian never-smoker females, among which EGFR mutations were the most frequently found mutation in lung adenocarcinomas of female never-smokers. Mutations in the TK domain of the EGFR were identified in those patients with refractory non-small cell lung cancer who achieved dramatic tumor responses to gefitinib. Although it is not

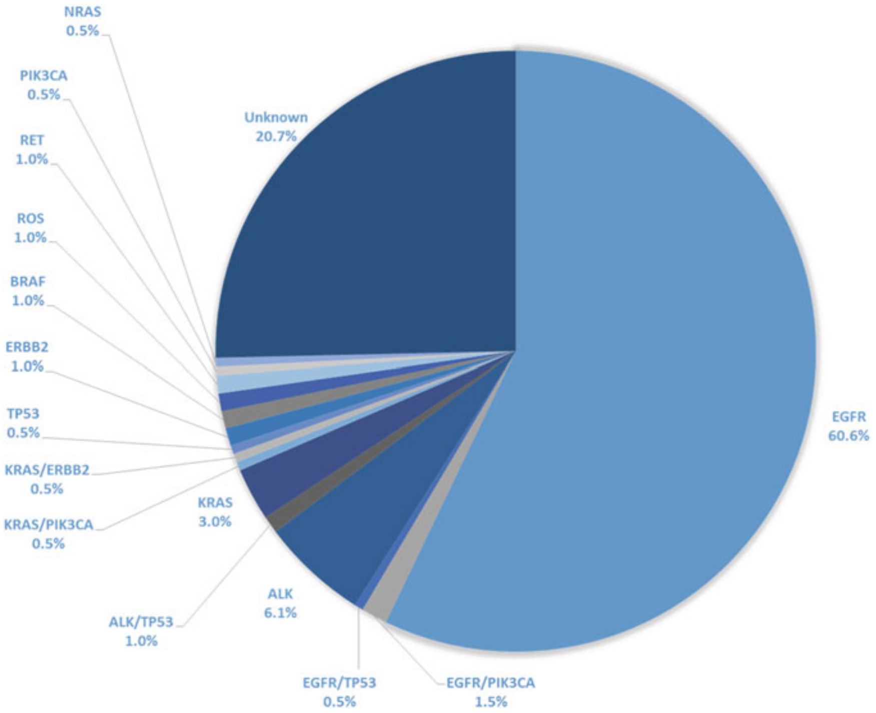


Fig. 8.2 Frequency of driver gene mutations in lung adenocarcinomas from East Asian never-smoker females (Ha et al. 2015)

well understood yet why NSCLCs develop in never-smokers, it has now been established that never-smoker-related NSCLCs comprise diseases distinct biologically to smoking-related NSCLCs: The former are characterized by the considerably high frequency in EGFR mutations, the latter show more frequent KRAS mutations and dominant unknowns (Oxnard et al. 2014; Sun et al. 2007). The major prevalence of driver oncogenes in lung adenocarcinomas from never-smokers certainly leads diseases putative oncogene-driven malignancies in these diseases, in which the treatment by kinase inhibitors highly benefits patients.

8.3 Clinical Proteogenomics

Millions of clinical samples are obtained every day for use in diagnostic tests that support clinical decision making. Clinical samples (tissues, biopsies, blood, *etc*.) can also be archived into repositories for use in future studies investigating the etiology of diseases using omics approaches. Therefore, infrastructure buildup of standardized biobanking is increasingly needed within the clinical omics community because the samples themselves have intrinsic values in the determination of 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 committees of medical institutes and hospitals. Many types of disease specimens exist, such as frozen and FFPE tissues; biopsies; and body fluids including blood, serum, plasma, and urine; interstitial fluid; cyst material; ascites fluid; and pancreatic juice.

8.3.1 Laser Microdisection and Protein Solubilization

In hospitals and medical institutes, tumor tissues obtained by surgical resection are typically fixed in 4 % paraformaldehyde and routinely processed

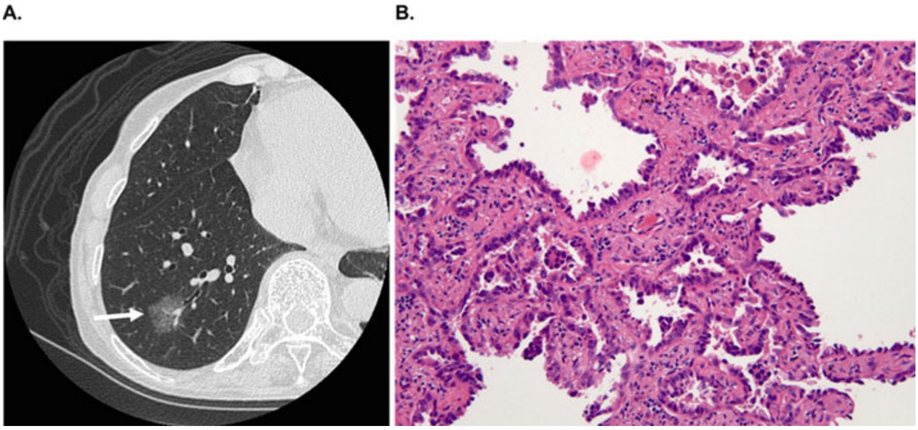


Fig. 8.3 (**a**) Focal ground-glass opacity on chest HRCT, which lesions are identified as AIS, MIA, or LEP. (**b**) A representative hematoxylin and eosin-stained image of LEP

for paraffin sectioning. Cancerous lesions can be identified on serial tissue sections stained with hematoxylin and eosin (HE). Figure 8.3 shows (A) focal ground-glass opacity on chest HRCT, which lesions are identified as AIS, MIA, or LEP and (B) a representative HE-stained image of LEP. Laser microdissection (LMD) makes it possible to collect target cells from a variety of FFPE cancer tissues. For shotgun proteomic analysis, 10-μm sections prepared from the same tissue block are attached onto DIRECTOR™ slides (OncoPlexDx, Rockville, MD, USA), deparaffinized twice with xylene for 5 min, rehydrated with graded ethanol solutions and distilled water, and stained by hematoxylin (Prieto et al. 2005; Hood et al. 2005; Kawamura et al. 2010; Nomura et al. 2011). Slides are air dried and subjected to LMD with a Leica LMD7000 (Leica Micro-systems GmbH, Ernst-Leitz-Strasse, Wetzlar, Germany). Typically, ca. 30,000 cells (ca. 8 mm^2) per tissue sample are transferred directly to a 1.5-mL low-binding plastic tube. Figure 8.4 exemplifies the hematoxylin-stained LEP tissue before and after LMD (C-1 and C-2, respectively).

Proteins/peptides from dissected cells can be extracted by following several protocols (Prieto et al. 2005; 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 at 95°C for 90 min, cooled on ice (3 min), and subsequently enzymatically digested, followed by reduction and alkylation. The liquid tissue digests can be stored at −20°C until proteomic analysis.

8.3.2 Protein Identification and Quantification

Recent advances in MS could make proteomics amenable to in-depth exploratory and targeted quantitative analysis of proteins expressed in a complex clinical specimen (Marko-Varga et al. 2011; Nakamura et al. 2012). MS is greatly advantageous due to its extremely high capability of capturing/identifying/sequencing of proteins/peptides expressed in a complex clinical specimen, with high sensitivity and high precision, unlike others (Nishimura and Tojo 2014). An exploratory proteomic analysis typically comprises extraction and/or direct tryptic digestion of all expressed proteins in a complex biological sample, and then the peptide mixture obtained is subjected to liquid chromatography (LC)/electrospray ionization-tandem MS“ShotGun” analytical platform so as to sequence these by searching against protein sequence databases. Protein identification in shotgun proteomic approaches (bottom-up) can be now performed by four peptide sequencing

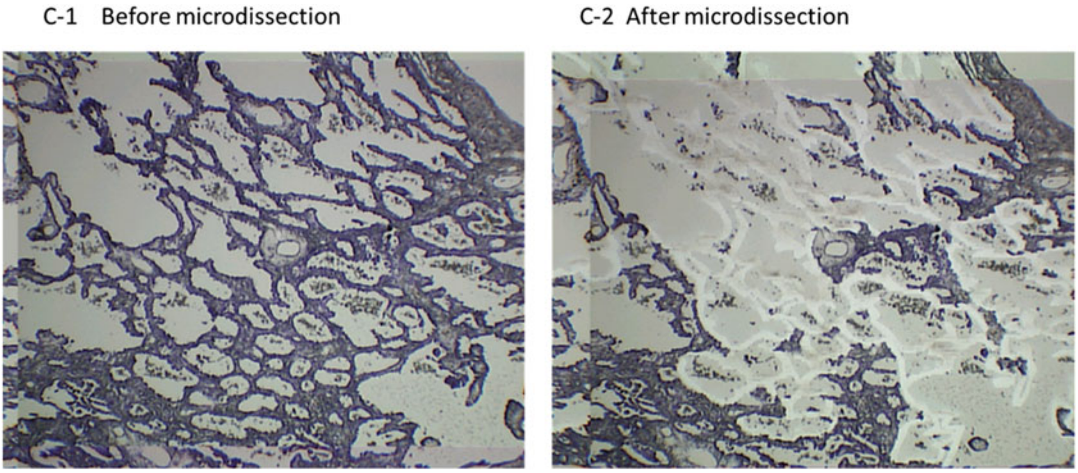


Fig. 8.4 The hematoxylin-stained LEP tissue before and after laser microdissection (C-1 and C-2, respectively). The DIRECTOR® slide is similar to a standard glass (uncharged) microscope slide but has an energy transfer coating on one side of the slide. Tissue sections are mounted on top of the energy transfer coating, and when the slide is turned over, the tissue faces down under the microdissection system. Analysis of targeting cells or tissue areas of interest is performed on the computer display. The laser energy is converted to kinetic energy upon striking the coating, vaporizing it, and instantly propelling selected tissue features into the collection tube

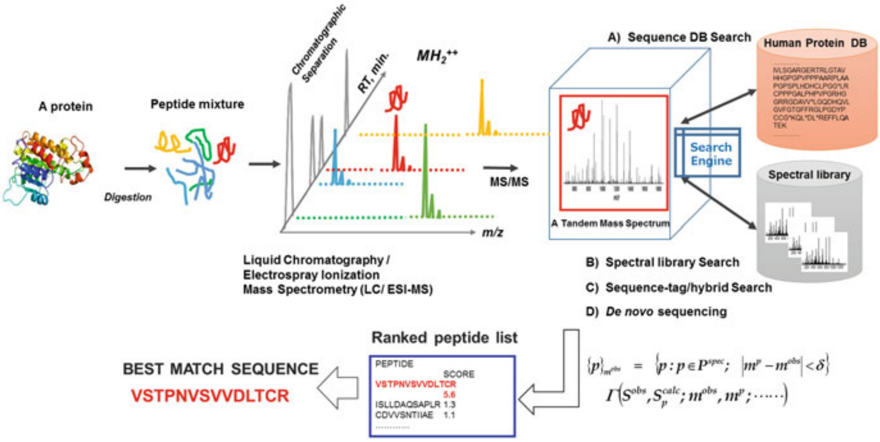

A protein
Digestion
Peptide mixture
Chromatographic Separation
RT, min.
MH_2^{++}
MS/MS
m/z
Liquid Chromatography / Electrospray Ionization Mass Spectrometry (LC/ ESI-MS)
A) Sequence DB Search
Human Protein DB
NLSGARGERTRLGTAV HHGPGPVPPPAARPLAA PGPSPLHDHCLPGG*LR CPPPGALPHPVPGRHG GRRGDAVV*LGQDHQVL GVFGTGFFRGLPGDYP CCG*KQL*DL*REFFLQA TEK
Search Engine
A Tandem Mass Spectrum
Spectral library
B) Spectral library Search
C) Sequence-tag/hybrid Search
D) *De novo* sequencing
Ranked peptide list
PEPTIDE
SCORE
VSTPNVSVVDLTCR 5.6
ISLLDAQSAPLR 1.3
CDVVSNTIIAE 1.1
BEST MATCH SEQUENCE
VSTPNVSVVDLTCR
$\{p\}_{m^{obs}} = \{p : p \in P^{spec};\ |m^p - m^{obs}| < \delta\}$
$\Gamma(S^{obs}, S_p^{calc}; m^{obs}, m^p; \cdots\cdots)$

Fig. 8.5 An illustration of exploratory mass spectrometry (MS)-based proteomic analysis. All expressed proteins in a complex biological sample are extracted and/or are subjected to direct tryptic digestion, and the peptide mixture obtained is subjected to liquid chromatography/electrospray ionization-tandem mass spectrometry "ShotGun" analytical platform so as to sequence these by queries against protein sequence databases

strategies using MS/MS spectra: (A) database search, (B) spectral library matching, (C) hybrid approaches using sequence-tag determination followed by database search, and (D) de novo sequencing as illustrated in Fig. 8.5. Several hundred to several thousand (more than 10,000 in some cases) different protein species can typically be identified in such exploratory clinical proteomic studies (Michalski et al. 2011; Panchaud et al. 2011; Geiger et al. 2010; Gillet et al. 2012; Bern et al. 2010; Physikron Mass Spectrometry Systems 2013; Gorshkov et al. 2015), in which label-free semi-quantitative comparison with statistical evaluation is mainly performed to elucidate proteins specifically relevant to a disease subtype.

8.4 Development of Proteogenomic Technologies to Identify/Quantify Mutated Proteins

8.4.1 Exploratory In-depth Analysis

Proteins are functionally dynamic since these biological molecules exhibit various forms of sequences, including not only post-translational modifications/truncations/variants but also mutations/rearrangements. For identification of such dynamic proteins from clinical samples, the development of proteome bioinformatics is of increasing importance. There are currently various gene–protein databases, which, however, remain to be consolidated. The total number of protein-coding genes is considered to be ~20,000, from which the canonical proteins are estimated to be ca. 68,000 in UniProt and ca. 38,400 in neXtProt. These databases have been most commonly used for protein identification in conventional proteomic studies. However, there are other databases not publicly available such as the MuPdb mutation database (ca. 1,200,000 species), the fusion-gene database (ca. 10,000 species), and the "Short" protein (the number of amino acids <100) database (ca. 6500 species), which have not been fully annotated yet. MuPdb contains all possible non-synonymous single nucleotide polymorphism (nsSNP) sequences known from genomic studies (Ensembl database).

MS-based proteomics generates high-quality sequential protein data on the human proteome. However, the lack of a comprehensive database including a complete collection of nsSNP products and fusion genes prevents identification of mutant and fusion proteoforms even if high-quality mass spectra are available. A recent study to determine the distribution of known oncogenic driver mutations in female never-smoker Asian patients with lung AC revealed that ca. 79 % of patients harbored driver gene mutations as previously described (Ha et al. 2015). Whereas multiplex gene analysis does not seem to be realistic, MS-based multiplex assays can be performed by querying proteome MS/MS datasets obtained from clinical samples against the variety of databases, which will highly likely identify new functional proteins relevant to therapeutic targets that have remained unknown until now (Fig. 8.6). Table 8.1 exemplifies a list of 35 proteins, identified with single amino acid polymorphisms, obtained from the proteomic dataset of patients with LEP lung cancer (single run, one sample) by querying against the mutated protein database, MuPdb.

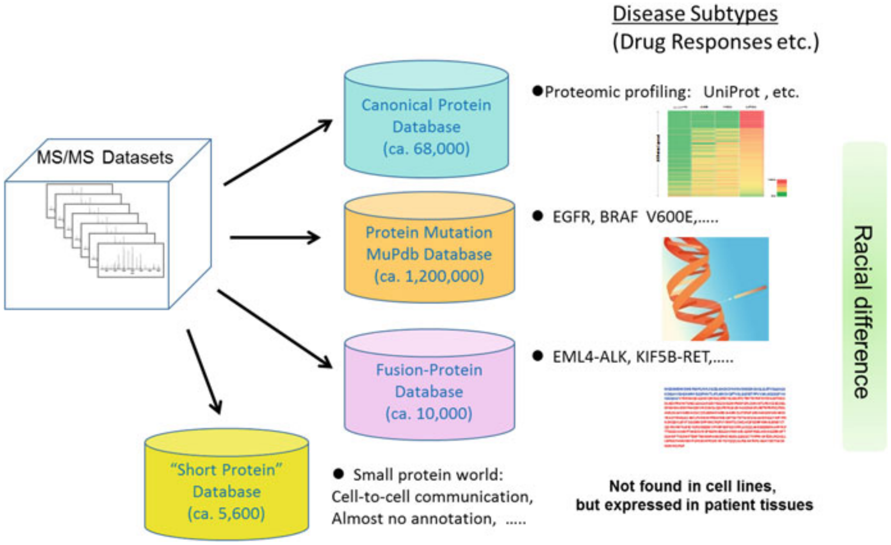

Disease Subtypes
(Drug Responses etc.)
MS/MS Datasets
Canonical Protein Database (ca. 68,000)
●Proteomic profiling: UniProt , etc.
Protein Mutation MuPdb Database (ca. 1,200,000)
● EGFR, BRAF V600E,.....
Fusion-Protein Database (ca. 10,000)
● EML4-ALK, KIF5B-RET,.....
"Short Protein" Database (ca. 5,600)
● Small protein world:
Cell-to-cell communication,
Almost no annotation, ...
Not found in cell lines,
but expressed in patient tissues
Racial difference

Fig. 8.6 Proteogenomic analysis of clinical proteome datasets by searching against various protein/gene databases: canonical UniProt and neXtProt protein db, MuPdb mutation db, fusion-gene db, and "short" protein (aa < 100) db

Table 8.1 The list of 35 proteins, identified with single aminoacid polymorphism (SAP), which were obtained from the proteomic dataset of LEP lung cancer (single LC-MS/MS run, one sample) by searching against MuPdb

No.	Protein name	Peptide	−10lgP	Mass	ppm	*m/z*	RT	Scan	Accession
1	HBD » Hemoglobin subunit delta	TAVN (+. 98) ALWGKVNVDEVGGEALGR	40.49	2255.1545	0.7	752.726	84.2	F1:9844	nx|NX_P02042-1-SNP-A-23-E|NX_P02042-1
2	COL1A1 » Collagen alpha-1(1) chain	SDKGETGEQGDR	56.44	1277.5483	−0.8	639.7809	6.65	F 1:421	nx|NX_P02452-1-SNP-G-1094-S|NX_P02452-1
3	COL1A1 » Collagen alpha-1(1) chain	GPSGPQGPGSPPGPK	38.46	1315.652	−0.6	658.8329	14.67	F1:1420	nx|NX_P02452-1-SNP-G-425-S|NX_P02452-1
4	COL1A1 » Collagen alpha-1(1) chain	GAAGEPSKAGER	39.46	1128.5522	0.1	565.2834	6.77	F 1:438	nx|NX_P02452-1-SNP-G-593-S|NX_P02452-1
5	COL1A1 » Collagen alpha-1(1) chain	GETGPAGPAGPVGPVGAR	62.87	1545.7899	1.1	773.9031	32.67	F2:3746	nx|NX_P02452-1-SNP-T-1075-A|NX_P02452-1
6	COL1A1 » Collagen alpha-1(1) chain	SGDRGETGPAGPAGPVGPVGAR	53.49	1960.9714	1.2	654.6652	29.34	F2:3319	nx|NX_P02452-1-SNP-T-1075-A|NX_P02452-1
7	FN1 » Fibronectin (FN)	GATYNIIVEALK	44.33	1290.7183	0.1	646.3665	75.94	F3:9125	nx|NX_P02751-14-SNP-V-2049-I|NX_P02751-14
8	ALB » Serum albumin	DAYKSEVAHR	42.79	1174.573	9.8	392.5355	6.73	F 1:433	nx|NX_P02768-1-SNP-H-27-Y|NX_P02768-1
9	COL1A2 » Collagen alpha-2(l) chain	GPPGESGASGPTGPIGSR	40.28	1579.759	0.8	790.8874	24.35	F3:2711	nx|NX_P08123-1-SNP-A-600-S|NX_P08123-1
10	COL1A2 » Collagen alpha-2(l) chain	GQAGLAGAR	46.87	799.43	0.2	400.7224	12.96	F1:1201	nx|NX_P08123-1-SNP-H-512-Q|NX_P08123-1
11	KRT19 » Keratin, type I cytoskeletal 19	FVSSSSSGGYGGGYGGVLTASDGLLAGNEK	46.14	2793.3093	2.1	932.1123	71.37	F1:8439	nx|NX_P08727-1-SNP-A-60-G|NX_P08727-1
12	COL6A2 » Collagen alpha-2(VI) chain	N(+.98) LQWIAGGTWTPSALK	34.26	1742.8992	4.1	872.4604	80.25	F1:9431	nx|NX_P12110-2-SNP-E-695-Q|NX_P12110-2
13	HIST1H1D » Histone H1.3	SLVN (+. 98) KGTLVQTK	40.37	1287.7397	−0.4	644.8769	44.65	F2:5261	nx|NX_P16402-1-SNP-S-90-N|NX_P16402-1
14	HLA-DPA1 » HLA class II histocompatibility antigen, DP alpha 1 chain	ETVWHLEEFGR	33.71	1401.6677	1.2	468.2304	62.29	F3:7508	nx|NX_P20036-1-SNP-Q-81-R|NX_P20036-1

Table 8.1 (continued)

No.	Protein name	Peptide	−10lgP	Mass	ppm	*m/z*	RT	Scan	Accession
15	HIST1H2BO » Histone H2B type 1-O	VMGIMN(+.98)SFVNDIFER	45.4	1771.8273	−4.6	886.9168	85.54	F3:10143	nx|NX_P23527-1-SNP-A-59-V|NX_P23527-1
16	TNC» Tenascin (TN)	VPGDQTSTIIR	51.85	1185.6354	−0.6	593.8246	31.48	F2:3595	nx|NX_P24821-2-SNP-Q-680-R|NX_P24821-2
17	PRSS3» Trypsin-3 [EC3.4.21.4]	LGEHN(+.98)INVLEGNEQFINAAK	46.02	2210.0967	0.5	737.7065	62.28	F3:7507	nx|NX_P35030-4-SNP-K-93-N|NX_P35030-4
18	CEACAM6 » Carcinoembryonic antigen-related cell adhesion molecule 6	SDPVTLNVLYGPDVPTISPSK	50.15	2198.147	2	1100.083	73.97	F1:8739	nx|NX_P40199-1-SNP-G-239-V|NX_P40199-1
19	HIST1H4A» Histone H4	RTVTAMDWYALK	52.75	1465.7963	−4.1	733.9024	76.1	F1:8989	nx|NX_P62805-1-SNP-K-80-R|NX_P62805-1
20	HIST1H4A» Histone H4	KTLTAMDWYALK	55.45	1451.8058	3	484.944	67.96	F1:8033	nx|NX_P62805-1-SNP-V-82-L|NX_P62805-1
21	ACTC1 » Actin, alpha cardiac muscle 1	DSYVGDEAQN (+. 98) KR	56.22	1381.611	0.5	691.8131	24.16	F1:2608	nx|NX_P68032-1-SNP-S-62-N|NX_P68032-1
22	EEF1A1 » Elongation factor 1-alpha 1 (EF 1 alpha 1)	LPLQEVYK	36.94	988.5593	0.7	495.2873	47.14	F1:5511	nx|NX_P68104-1-SNP-D-252-E|NX_P68104-1
23	ACTA1 » Min, alpha skeletal muscle	KVLYANNVMSGGTTM(+15.99)YPGIAD p	34.78	2373.1458	6.2	792.0607	56.37	F3:6778	nx|NX_P68133-1-SNP-D-294-V|NX_P68133-1
24	HBB » Hemoglobin subunit beta	FFESFGDLSTPDTVMGNPK	39.16	2087.9509	1.8	1044.9846	69.67	F3:8391	nx|NX_P68871-1-SNP-A-54-T|NX_P68871-1
25	HBB » Hemoglobin subunit beta	FFQ(+.98)SFGDLSTPDAWI(+15.99)GNPK	39.82	2073.9353	2.9	1037.9779	69.23	F1:8186	nx|NX_P68871-1-SNP-E-44-Q|NX_P68871-1
26	HBB » Hemoglobin subunit beta	SAVTALWAK	47.12	945.5283	0.7	473.7718	56.1	F2:6693	nx|NX_P68871-1-SNP-G-17-A|NX_P68871-1

27	HSPG2 » Basement membrane-specific heparan sulfate proteoglycan core protein (HSPG)	SIQYSPQ(+ .98) LEDAGS R	38.26	1550.7212	–1.2	776.3669	41.83	F3:4936	nx|NX_P98160-1-SNP-E-95-Q|NX_P98160-1
28	TUBA3C » Tubulin alpha-3C/D chain	AVFVDLEPWLDEVR	50.05	1700.8984	1.6	851.4579	78.2	F3:9377	nx|NX_Q13748-1-SNP-V-75-L|NX_Q13748-1
29	QPRT » Nicotinate-nucleotide pyrophosphoiylase [carboxyl ating] [EC 2.4.2.19]	YGLLVGGAVSHR	50.24	1227.6724	1.5	410.232	43.34	F2:5096	nx|NX_Q15274-1-SNP-A158-V|NX_Q15274-1
30	TUBB8 » Tubulin beta-8 chain	PVLVDLEPGTM DSVR	38.7	1626.8286	1.3	814.4226	71.14	F1:8412	nx|NX_Q3ZCM7-1-SNP-A-63-P|NX_Q3ZCM7-1
31	ACTBL2 » Beta-actin-like protein 2	VAPDEHPILLTEAPPN(+.98) P KIN (+. 98)	29.38	2322.2219	4.9	775.085	56.63	F2:6758	nx|NX_Q562R1-1-SNP-L-111-P|NX_Q562R1-1
32	TUBA3E » Tubulin alpha-3E chain	AVFVDLEPTWEEVR	41.92	1700.8984	–0.9	851.4557	78.18	F2:9249	nx|NX_Q6PEY2-1-SNP-D-76-E|NX_Q6PEY2-1
33	POTEE » POTE ankyrin domain family member E	SYQLPDGQ(+.98)VITIGNER	39.83	1789.8846	1.7	895.9511	59.26	F3:7130	nx|NX_Q6S8J 3-1-SNP-E-941-Q|NX_Q6S8J 3-1
34	HIST3H2BB » Histone H2B type 3-B	SMGIM(+15.99) NSFVNDIFER	48.72	1774.8019	0.8	888.4089	77.77	F1:9176	nx|NX_Q8N257-1-SNP-A-59-S|NX_Q8N257-1
35	GCN1L1» Translational activator GCN1 (HsGCN1)	IIIEDLLEATR	39.02	1284.7289	5.1	643.375	82.41	F1:9661	nx|NX_Q92616-1-SNP-Y-2155-D|NX_Q92616-1

*(+0.98) and (+15.99) indicate deamidation on asparagine and oxidation, respectively

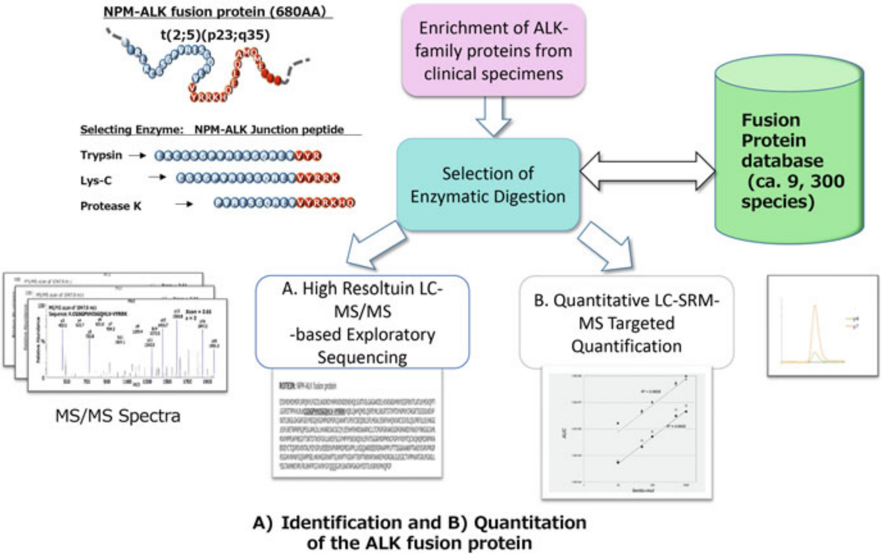

NPM-ALK fusion protein (680AA)
t(2;5)(p23;q35)
Selecting Enzyme: NPM-ALK Junction peptide
Trypsin
Lys-C
Protease K
Enrichment of ALK-family proteins from clinical specimens
Selection of Enzymatic Digestion
Fusion Protein database (ca. 9, 300 species)
A. High Resoltuin LC-MS/MS -based Exploratory Sequencing
B. Quantitative LC-SRM-MS Targeted Quantification
MS/MS Spectra
A) Identification and B) Quantitation of the ALK fusion protein

Fig. 8.7 A workflow of identifying and quantifying fusion gene products, in which the NPM-ALK fusion protein (680 aa) produced by the gene translocation t(2;5)(p23;q35) is shown as an example

8.4.2 Quantitative Identification of Oncogenic Fusion Gene Products

Gene fusion events represent the most common type of genomic rearrangement resulting from inversions, interstitial deletion, or translocations. The first fusion gene, BRC-ABL1was discovered by Peter Nowell & David Hungerford (1960) and has been a diagnostic marker for chronic myelogenous leukemia as well as the drug target of imatinib (Gleevec™). The discovery of TMPRSS2-ERG fusions in prostate cancer (Tomlins et al. 2008) and EML4-ALK fusion in NSCLC tumors (Soda et al. 2007) suggested a relatively frequent occurrence of gene fusion events in solid tumors, which generate novel oncogenic fusion proteins. These fusion gene products are mainly restricted to tumor cells, which would be useful diagnostic and therapeutic targets.

The fusion protein FusPdb database, built from the integrated chimeric gene database, ChiTaRS (2013), currently contains ca. 9300 candidates. Figure 8.7 illustrates the workflow of identifying and quantifying fusion gene products, in which the NPM-ALK fusion protein (680 aa) produced by the gene translocation t(2;5)(p23;q35) is showed as an example. Recent advances in mass spectrometric make it possible to identify numerous fusion junction peptides/fusion proteins by LC–MS/MS in high-resolution and high sensitivity modes, in which resulting datasets are queried against the FusPdb/canonical protein databases and/or directly against the ChiTaRS database using appropriate searching software, such as COMET (Eng et al. 2015). Prior to MS-based analysis of fusion gene products of interest, it is key to develop methodologies for affinity-based enrichment from a complex clinical sample, followed by efficient digestion. Junction peptides can be designed using the FuseProt database, which are all possible peptides containing the junction position of a fusion protein. For MS-based quantification of targeted fusion-proteins of interest, unique junction peptides involving junction points are designed by selecting appropriate enzymatic proteases, in which their enzymatic miss-cleavages are also taken into account. MS-based targeted assays would typically be performed using the acquisition modes of the single-ion monitoring (SIM) and selected-reaction monitoring (SRM), where absolute quantification can be achieved with spiked authentic stable isotope-labeled junction peptide (AQUA junction peptide) of a known amount. There are the major three variants of EML4-ALK fusion gene products. Figure 8.8 shows (A) the 11 SRM-transition peaks obtained for the authentic EML4-ALK variant-1 junction peptide, k.YIMSNSGDYEIYL-YRRK.h, being produced via Lys-C digestion and (B) the SIM quantitation of the authentic EML4-ALK Variant-3 junction peptide, k.NSQ-VYRRK.h (a Lys-C digestion product) of ca. 10–1000 fmol spiked into a plasma matrix (ca. 100 ng/μl). Both were measured on a Q-Exactive high-resolution mass spectrometer. The MS-based targeted quantification assay of multiple fusion gene products including their variants is in principle accurate and flexible, by which multiple targeted fusion gene products can be quantitatively monitored simultaneously with tissue/biopsy of a patient. For example, simultaneous MS-based quantitation of both EML4-ALK and KIF5B-RET—a multiplex assay—can be designed as follows: (1) targeted fusion proteins (including variants) can be enriched using appropriate antibodies, (2) respective enriched samples are digested by specific proteases, and (3) the respective peptide-mixture samples are combined into one sample, which is subjected to targeted MS-based quantification. A MS-based multiplex fusion protein assay has a high potential to facilitate establishment of the definitive diagnosis, benefiting the patient by enabling administration of optimal

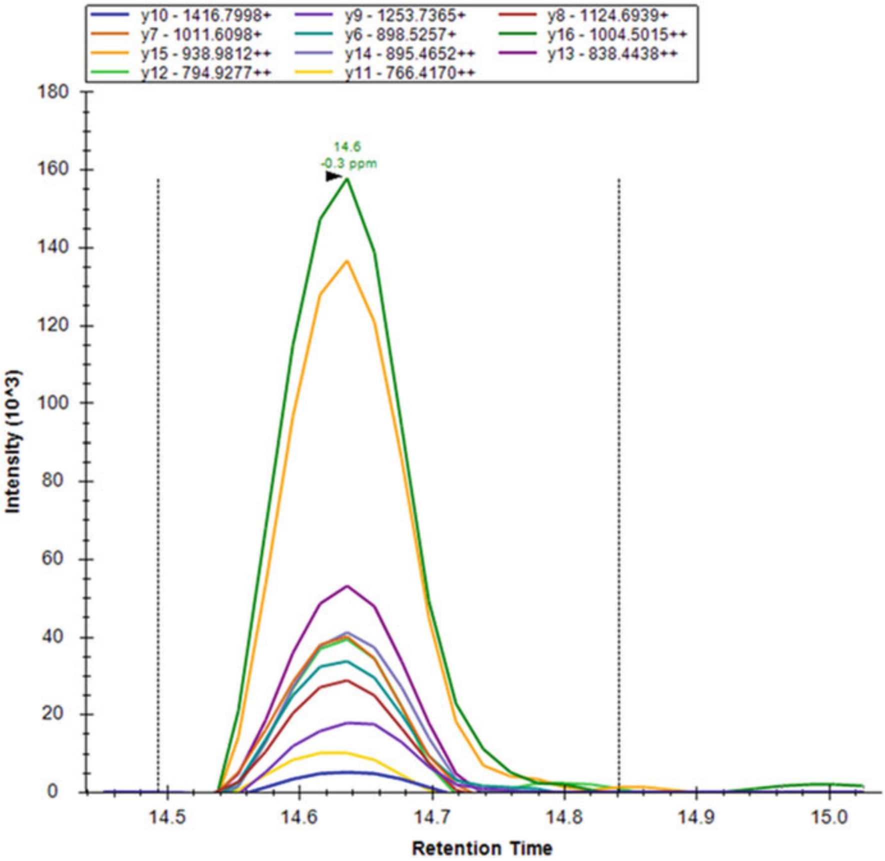


Fig. 8.8 (**a**) The 11 SRM-transition peaks obtained for the authentic EML4-ALK variant-1 junction peptide, k.YIMSNSGDYEIYL-YRRK.h, being produced via Lys-C digestion. (**b**) SIM quantitation for the authentic EML4-ALK variant-3 junction peptide, k.NSQ-VYRRK.h (a Lys-C digestion product) of ca. 10–1000 fmol spiked into a plasma matrix (ca. 100 ng/μl)

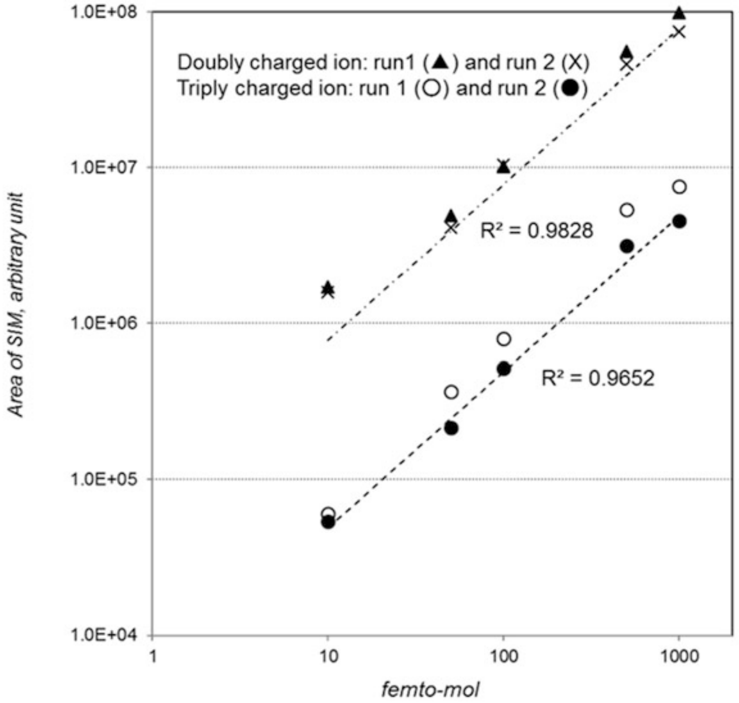


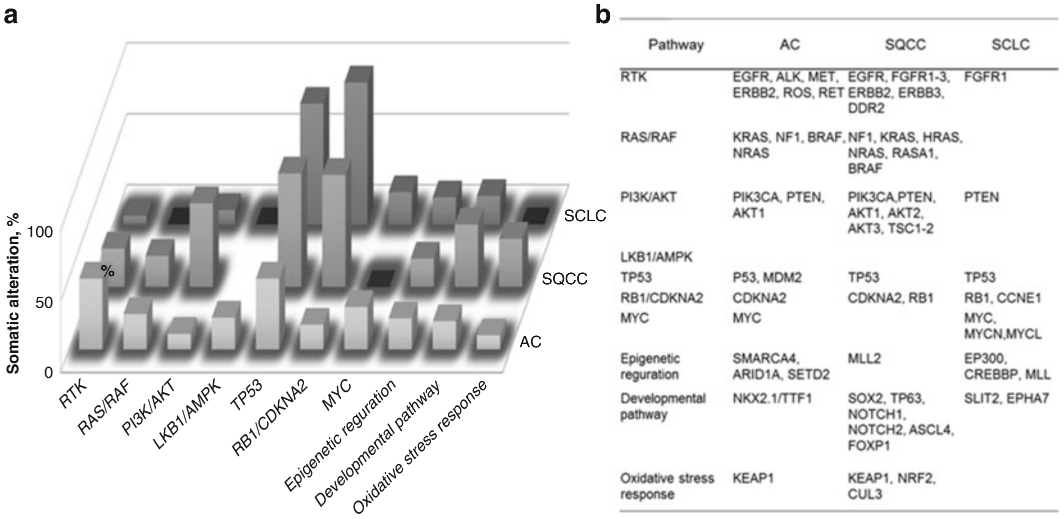
Fig. 8.8 (continued)

treatment, and comes with a low assay cost, although further development of the technology is necessary.

8.5 Cellular Pathways Affected by Somatic Mutations

The identification of recurrent mutations in EGFR and fusions involving ALK and other receptor tyrosine kinases has greatly transformed the standard of treatment of patients with lung AC. Current guidelines recommended the molecular genotyping of AC to routinely include the EGFR and ALK status, alterations which are found to exist in ca. 25 % of patients with AC who benefit more from approved targeted inhibitor therapies than from conventional chemotherapy. Such somatic alterations, mutations, and fusions in lung cancers frequently affect cellular pathway activities involved in lung cancer subtypes. Figure 8.9 summarizes cellular pathways, the activities of which are affected by somatic alterations in lung cancer subtypes, namely AC, squamous cell carcinoma, and SCLC (Shtivelman et al. 2014).

The most important information is how proteins expressed significantly in a disease subtype interplay with other key proteins and pathways to evaluate biomarker candidates and therapeutic targets. Several open PPI databases are available; current versions include Reactome (Reactome Pathway database 2016) and BioGRID (The Biological General Repository for Interaction Datasets 2016), and PPI network analysis can be performed by designated network construction algorithms, using, for example, the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (STRING 10 2016) and the Cytoscape (Cytoscape3 2014), a software environment for integrated models of biomolecular interaction networks. PPI networks elucidated so far consist of nodes and edges, where nodes are proteins experimentally identified, and edges are the predicted functional associations based on primary databases comprising KEGG and GO (2016), the primary literature, and so on. Thus, it has become possible to elucidate protein networks



Pathway	AC	SQCC	SCLC
RTK	EGFR, ALK, MET, ERBB2, ROS, RET	EGFR, FGFR1-3, ERBB2, ERBB3, DDR2	FGFR1
RAS/RAF	KRAS, NF1, BRAF, NRAS	NF1, KRAS, HRAS, NRAS, RASA1, BRAF	
PI3K/AKT	PIK3CA, PTEN, AKT1	PIK3CA,PTEN, AKT1, AKT2, AKT3, TSC1-2	PTEN
LKB1/AMPK			
TP53	P53, MDM2	TP53	TP53
RB1/CDKNA2	CDKNA2	CDKNA2, RB1	RB1, CCNE1
MYC	MYC		MYC, MYCN,MYCL
Epigenetic reguration	SMARCA4, ARID1A, SETD2	MLL2	EP300, CREBBP, MLL
Developmental pathway	NKX2.1/TTF1	SOX2, TP63, NOTCH1, NOTCH2, ASCL4, FOXP1	SLIT2, EPHA7
Oxidative stress response	KEAP1	KEAP1, NRF2, CUL3	

Fig. 8.9 (**a**) Cellular pathways affected by somatic alterations and (**b**) genes involved in each type of lung cancer are listed in order of the frequency of alterations found. *AC* adenocarcinoma, *SCLC* small-cell lung carcinoma, *SQCC* squamous cell carcinoma (Shtivelman et al. 2014)

relevant to a disease subtype using its proteomic datasets. The concept of biomarkers has been changing from conventional biomarkers—single proteins—to specific protein networks or dynamically varying networks (Wu et al. 2014), since diseases can be regarded also as dynamic network disorders. PPI networks dynamically activated or deactivated in a disease subtype of interest would be directly associated with its responsible molecular mechanisms, which would lead to discovery of therapeutic and drugable targets.

Figure 8.10 shows the STRING PPI networks developed from the LEP proteome datasets obtained from a clinical proteomic analytical study of cancerous cells laser microdissected from lung cancer FFPE-tissue specimens (Kato et al. 2015). The ErbB signaling pathway is associated with several cancer pathways. The ErbB family represents epidermal growth factor receptors, which play an important role in tumor growth. Overexpression of EGFR occurs in around 60 % NSCLCs, with patients with AC having the highest frequency (Shtivelman et al. 2014). Hypoxia-inducible factors (HIFs) regulate the transcription of genes that mediate the response to hypoxia (reduced O_2 availability) (Semenza 2010). Diverse products of HIF-1 action such as induction of the Met protein, hepatocyte growth factor, followed by Met receptor activation, may result in the poor prognosis associated with hypoxic tumors, which are indeed more aggressive than their well-oxygenated counterparts. Molecules participating in the ErbB and HIF-1 signaling pathways are denoted by orange and red circles in Fig. 8.10, respectively (Kato et al. 2015). Figure 8.11 illustrates the results of the STRING gene set enrichments for LEP, MIA, and AIS obtained for the 24 cancer-related KEGG pathways, which were elucidated with their significance rank $p < 0.05$ after correction by false discovery rate (FDR) (Kato et al. 2015). It was revealed how functional participation of expressed proteins alters dramatically throughout disease stages, reflecting the mechanisms of disease progression. Thus, MS-based exploratory proteomics utilizing clinical specimens is a promising analytical platform, which makes it possible to reveal molecular networks relevant to a disease subgroup, drug responders or nonresponders, good or poor prognosis, drug resistance, and so on.

It should be emphasized that both somatic mutations and cellular pathways in disease subtypes are mutually strongly connected, and so both are needed to be unveiled to understand molecular mechanisms of a disease subtype.

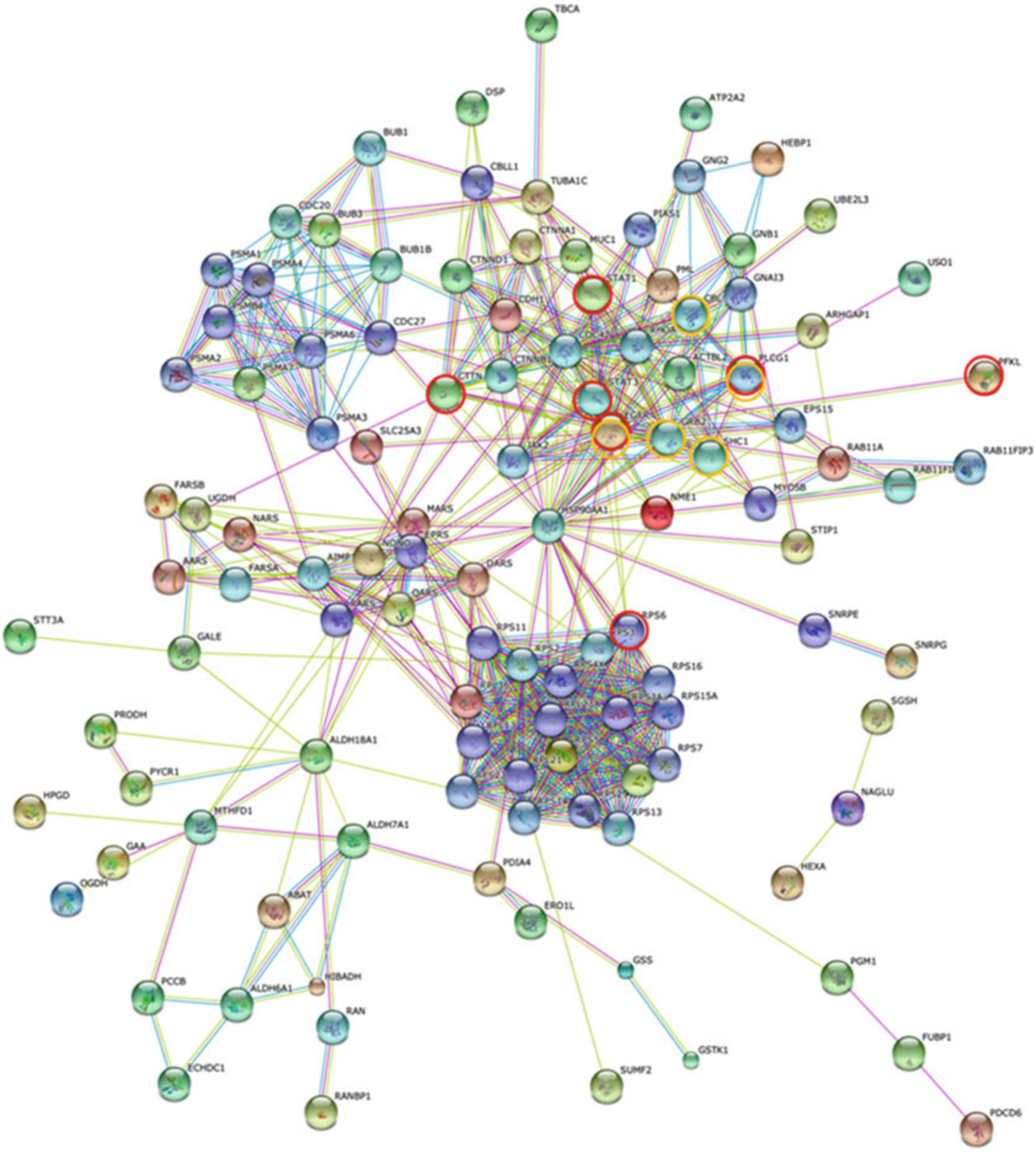


Fig. 8.10 The STRING protein–protein interaction networks developed from the LEP proteome datasets obtained from a clinical proteomic analytical study of the cancerous cells laser-microdissected from lung cancer formalin-fixed paraffin-embedded-tissue specimens

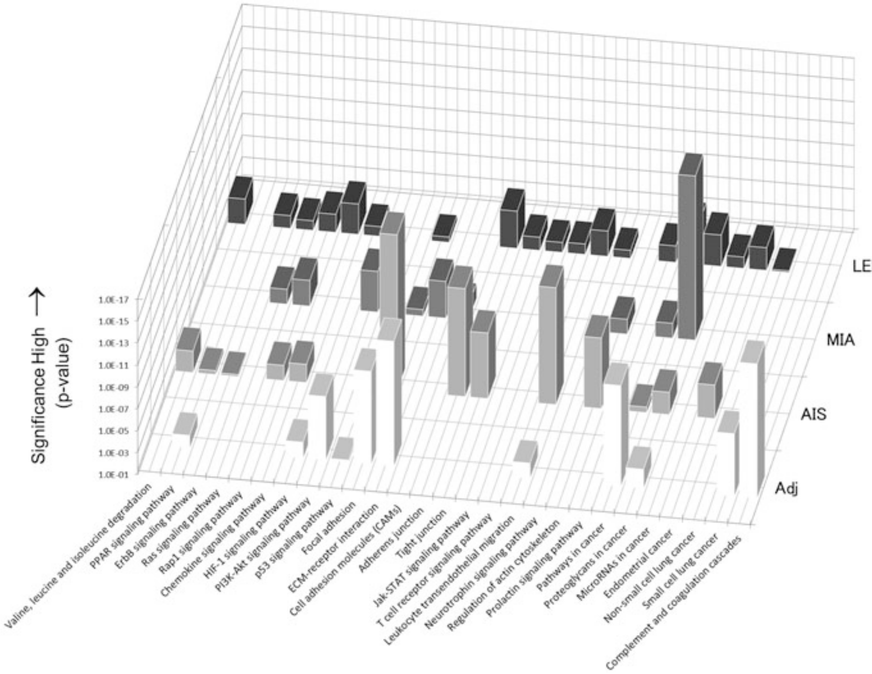


Fig. 8.11 Results of the STRING gene set enrichments for LEP, MIA, and AIS obtained for the 24 cancer-related KEGG pathways (significance rank $p<0.05$ after correction by false discovery rate), which revealed how functional participations of expressed proteins alter dramatically throughout disease stages, reflecting mechanisms of disease progression (Kato et al. 2015)

8.6 Drug Resistance and Gatekeeper Mutations

8.6.1 Third-Generation EGFR Tyrosine Kinase Inhibitor (TKI) Drugs

First-generation EGFR-TKIs include Iressa™ (gefitinib) and Tarceva™ (erlotinib), which target activated EGFR mutations, including the L858R point mutation in Exon 21 and small deletions around E746-A750 in Exon 19, in the tyrosine kinase domain. Such mutations are frequently seen in lung adenocarcinomas (ACs) in East Asian populations. These two drugs show remarkable therapeutic effects in 40–50 % of patients with lung cancer with AC in China, Korea, and Japan (Mitsudomi et al. 2010). However, drug resistance has been reported to be associated with the administration of first-generation EGFR-TKIs, where an acquisition of the new T790M EGFR mutation in patients with EGFR mutation-positive lung cancer is considered the most frequent cause of resistance. Later, the second-generation EGFR-TKI Gilotrif™ (afatinib) (Hirsh 2015; Bennouna and Moreno Vera 2015) was introduced; however, it appears that this TKI is no longer able to be bound due to the three-dimensional structural change of EGFR caused by the T790M mutation. Recently, the third-generation EGFR-TKIs, AZD9291 (Jänne et al. 2015) and CO-1686 (rociletinib) (Sequist et al. 2015), demonstrated a high antitumor effect on non-small-cell lung cancer (NSCLC), which had become resistant to first-generation EGFR-TKIs; these TKIs have been found to selectively and irreversibly inhibit both EGFR with TKI-activated mutations and T790M and are currently in the development phase.

8.6.2 EGFR-TKI Drug Resistance and Dynamic Variation of Mutations

Recently, it was reported that AZD9291 showed high antitumor activity in EGFR-T790M-positive patients with advanced NSCLC whose disease progressed with previous EGFR-TKI treatment and that rociletinib, which exhibited activity in both T790M (+) and (−) patients in a preclinical model, benefits EGFR-mutation-positive patients with EGFR-T790M.

The third-generation EGFR-TKI AZD9291 demonstrated a notable result in that the median PFS was 13.1 months in a phase 2 trial limited to T790M-positive cases (AURA trial, NCT01802632: AURA 2015). No difference was observed in the response rate between patients with Del19/T790M and L858R/T790M, which suggested that AZD9291 is specific to T790M (AURA 2015). The current diagnosis using gene analysis only judges whether the T790M mutation is present or not, and it is highly possible that a clone of T790M exists from the beginning. Regarding tumor heterogeneity, three different scenarios of tumors carrying two EGFR mutations have been contemplated, in which activating and resistant mutations exist: (1) in cis on the same allele, (2) in trans on different alleles, or (3) in different clones (Fig. 8.12) (Leone 2013).

Therefore, it is critical to develop a precise methodology for T790M quantification, by which a correlation between the extent of T790M expression and drug efficacy can be investigated, thereby making it possible to define a cutoff value for T790M abundance. Figure 8.13 shows EGFR mutations in NSCLC, the crystal structure of the kinase domain of EGFR in complex with gefitinib (based on Protein Data Bank accession code 2ITY), and the location of the EGFR mutations (RCSB 2015).

The efficacy of AZD9291 would also be limited because resistance will be acquired to this drug. C797 codon mutations in the EGFR tyrosine kinase-binding sites have been reported as a resistance mechanism for AZD9291 in T790M. A recent study suggested that there are both T790M (+) and (−) clones at baseline and that AZD9291 intervenes with the resistance of T790M (−) cells, although AZD9291 effectively suppresses the growth of T790M (+) cells, which might be bypassed due to the activation of the HER2 and/or MET pathway. Furthermore, EGFR-TKI drug resistance has been associated with a new C797S mutation found near T790M (Thress et al., 2015) There are two plausible patterns involved in AZD9291 resistance: (1) both T790M (+) and C797S (+) or (2) T790M (−) with bypassing occurring in a different signaling pathway such as HER2 or MET. Both resistance patterns might coexist within one type of tissue and within both primary and metastatic tumors. It has been reported that gefitinib might be effective for C797S (+) cases (Ercan et al. 2015).

8.7 Perspective

Both identification and quantification of EGFR gatekeeper mutations, including mutation heterogeneity within a tissue, need to be performed using lung cancer tissue specimens obtained from patients to improve the treatment for patients with EGFR mutation-positive NSCLC, including:

1. Identification and quantitation data of targeted EGFR mutated proteins
2. Exploratory mass spectrometry (MS)-based clinical proteogenomic analysis of mutated proteins, including investigation of mutation heterogeneity within a tissue
3. Analysis of dynamic protein–protein interaction (PPI) networks of proteins significantly related to a subgroup of patients with lung cancer with acquired resistance

The most interesting investigation involves the pairwise MS-based proteogenomic analysis of EGFR mutation-positive (frozen and/or FFPE) tissues of lung cancer obtained prior to EGFR-TKI treatment and after acquisition of the drug resistance, which will unveil detailed and direct molecular information on EGFR-TKI resistance.

In Japan, high-quality clinical specimens with detailed clinical and pathological information

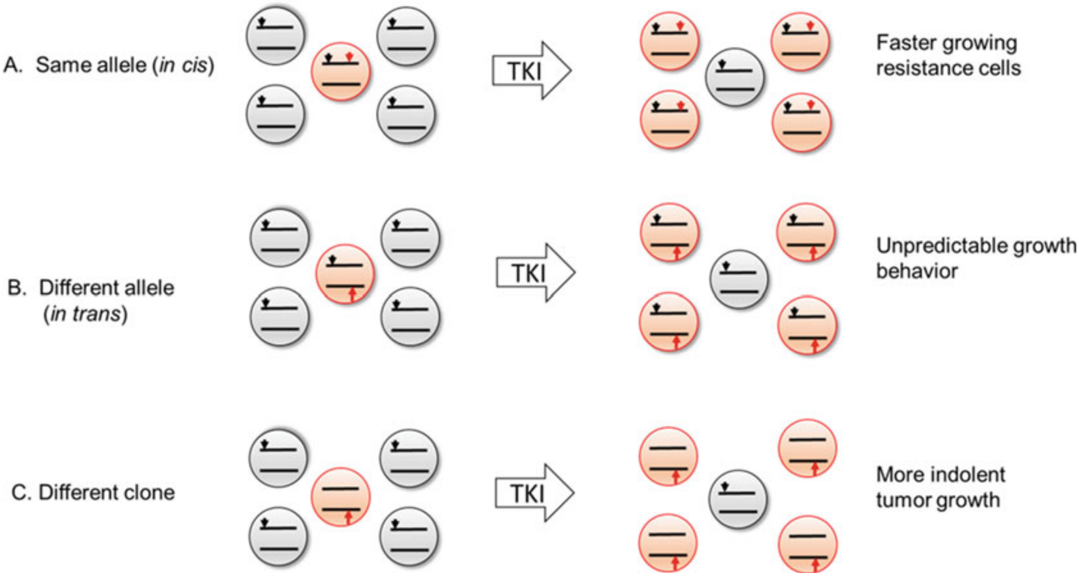

A. Same allele (*in cis*)
TKI
Faster growing resistance cells
B. Different allele (*in trans*)
TKI
Unpredictable growth behavior
C. Different clone
TKI
More indolent tumor growth

Fig. 8.12 Three different scenarios of tumors carrying two EGFR mutations have been suggested, including activating and resistant mutations: (1) *in cis* on the same allele, (2) *in trans* on different alleles, and (3) in different clones. *Black circles*, cells with EGFR-activating mutation; *red circles*, cells with activating or resistant mutation; *black lines*, different alleles; *black arrowhead*, activating mutation; *red arrowhead*, resistant mutation. *TKI* tyrosine kinase inhibitor (Leone 2013)

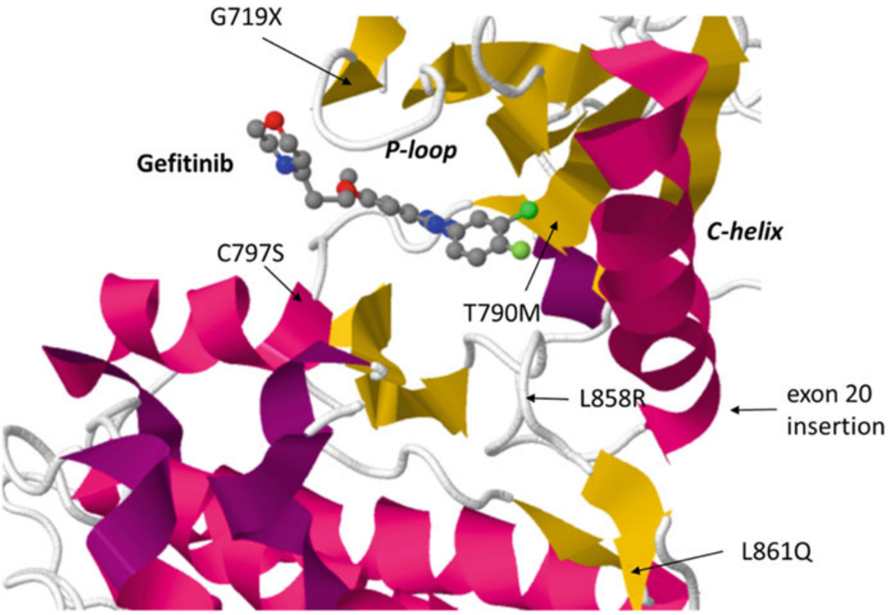


Fig. 8.13 EGFR mutations in nonsmall-cell lung cancer, the crystal structure of the kinase domain of EGFR in complex with gefitinib (based on Protein Data Bank accession code 2ITY), and the location of the EGFR mutations (RCSB Protein Data Bank; Thress et al. 2015)

have been archived for several years within medical institutes and hospitals and include even early-stage cancers. In this decade, it was revealed that drug efficacy differs by race (*e.g.*, Caucasians vs. Asians) due to oncogenic driver mutations / fusions specific to each race, exemplified by gefitinib and erlotinib. In contrast to time-consuming genomic analysis, MS-based proteogenomic approaches enable direct analysis of mutated and fusion proteins expressed in a clinical sample, which will provide a powerful solution for the stratification of patients and drug discovery (Precision Medicine). When a distinct clinical study design using valuable clinical samples, so as to say national assets, is established and performed by teaming up scrupulously with clinicians, an innovative treatment and drug discovery pipeline can be delivered from Japan, opening a gateway to Asia and its population of 3.9 billion.

Acknowledgments TN and HN thank Professor Emeritus Dr. Harubumi Kato at Tokyo Medical University for his valuable guidance regarding to clinical study designs, and Drs. Melinda Rezeli, Charlotte Welinder, Johan Malm, the late Thomas E. Fehniger, and György Marko-Varga at the Center of Excellence in Biological and Medical Mass Spectrometry, Biomedical Center, Lund University for their collaborative developments of MS-based technologies. TN especially thanks Dr. Ákos Végvári at Pharmacology & Toxicology, University of Texas Medical Branch, Galveston TX USA, for his instruction to MS-based protein-mutation studies by utilizing the MuPdb.

References

AURA. (2015). *National Institutes of Health. AZD9291 First time in patients ascending dose study (AURA)*. 2015: Available from https://clinicaltrials.gov/ct2/show/NCT01802632. 16 Jan 2016.

Battafarano, R. J., Fernandez, F. G., Ritter, J., Meyers, B. F., Guthrie, T. J., Cooper, J. D., & Patterson, G. A. (2005). Large cell neuroendocrine carcinoma: an aggressive form of non-small cell lung cancer. *Journal of Thoracic and Cardiovascular Surgery, 130*, 166–172.

Bennouna, J., Moreno Vera, S. R. (2015). *Afatinib-based combination regimens for the treatment of solid tumors: rationale, emerging strategies and recent progress*. Future Oncology, 2015 Nov 25. doi:10.2217/fon.15.310. Available from http://www.futuremedicine.com/doi/abs/10.2217/fon.15.310. 16 Jan 2016.

Bern, M., Finney, G., Hoopmann, M. R., Merrihew, G., Toth, M. J., & MacCoss, M. J. (2010). Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Analytical Chemistry, 82*, 833–841.

BioGRID-The Biological General Repository for Interaction Datasets. Available from http://thebiogrid.org/. 15 Jan 2016.

Camidge, D. R., & Doebele, R. C. (2012). Treating ALK-positive lung cancer-early successes and future challenges. *Nature Reviews Clinical Oncology, 9*(5), 268–277.

ChiTaRS (the database of chimeric transcripts and RNA-Sequencing data database). (2013). Available from http://chitars.bioinfo.cnio.es/. 16 Jan 2016.

Cytoscape – An open source software platform for visualizing complex networks and integrating these with any type of attribute data. Available from http://www.cytoscape.org/. 15 Jan 2016.

Dresler, C. M., Ritter, J. H., Patterson, G. A., Ross, E., Bailey, M. S., & Wick, M. R. (1997). Clinical pathologic analysis of 40 patients with large cell neuroendocrine carcinoma of the lung. *The Annals of Thoracic Surgery, 63*, 180–185.

Eng, J. K., Hoopmann, M. R., Jahan, T. A., Egertson, J. D., Noble, W. S., & MacCoss, M. J. (2015). A deeper look into comet—Implementation and features. *Journal of the American Society for Mass Spectrometry, 26*, 1865–1874.

Ercan, D., Choi, H. G., Yun, C. H., Capelletti, M., Xie, T., Eck, M. J., Gray, N. S., & Jänne, P. A. (2015). EGFR mutations and resistance to irreversible pyrimidine-Based EGFR Inhibitors. *Clinical Cancer Research, 21*(17), 3913–3923.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer, 127*, 2893–2917.

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

Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., & Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics, 11*, 1–17.

Gorshkov, V., Verano-Braga, T., & Kjeldsen, F. (2015). SuperQuant: A data processing approach to increase Quantitative Proteome Coverage. *Analytical Chemistry, 87*, 6319–6327.

Gu, J., Lu, C., Guo, J., Chen, L., Chu, Y., Ji, Y., & Ge, D. (2013). Prognostic significance of the IASLC/ ATS/ ERS classification in Chinese patients-A single institution retrospective study of 292 lung adenocarcinoma. *Journal of Surgical Oncology, 107*(5), 474–480.

Ha, S. Y., Choi, S. J., Cho, J. H., Choi, H. J., Lee, J., Jung, K., Irwin, D., Liu, X., Lira, M. E., Mao, M., Kim, H. K., Choi, Y. S., Shim, Y. M., Park, W. Y., Choi, Y. L., & Kim, J. (2015). Lung cancer in never-smoker Asian females is driven by oncogenic mutations, most often involving EGFR. *Oncotarget, 6*(7), 5465–5474.

Hirsh, V. (2015). Next-generation covalent irreversible Kinase inhibitors in NSCLC: Focus on Afatinib. *BioDrugs, 29*(3), 167–183.

Hood, B. L., Darfer, M. M., Guiel, T. G., Furusato, B., Lucas, D. A., Ringeisen, B. R., Sesterhenn, I. A., Conrads, T. P., Veenstra, T. D., & Krizman, D. B. (2005). Proteomic analysis of formalin-fixed prostate cancer tissue. *Molecular & Cellular Proteomics, 4*, 1741–1753.

Hu, H., Pan, Y., Li, Y., Wang, L., Wang, R., Zhang, Y., Li, H., Ye, T., Zhang, Y., Luo, X., Shao, L., Sun, Z., Cai, D., Xu, J., Lu, Q., Deng, Y., Shen, L., Ji, H., Sun, Y., & Chen, H. (2014). Oncogenic mutations are associated with histological subtypes but do not have an independent prognostic value in lung adenocarcinoma. *Journal of OncoTargets and Therapy, 7*, 1423–1437.

Jänne, P. A., Yang, J. C., Kim, D. W., Planchard, D., Ohe, Y., Ramalingam, S. S., Ahn, M. J., Kim, S. W., Su, W. C., Horn, L., Haggstrom, D., Felip, E., Kim, J. H., Frewer, P., Cantarini, M., Brown, K. H., Dickinson, P. A., Ghiorghiu, S., & Ranson, M. (2015). AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer. *The New England Journal of Medicine, 372*(18), 1689–1699.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians, 61*, 69–90.

Kato, Y., Nakamura, H., Tojo, H., Nomura, M., Nagao, T., Kawamura, T., Kodama, T., Ohira, T., Ikeda, N., Fehniger, T., Marko-Varga, G., Nishimura, T., & Kato, H. (2015). A proteomic profiling of laser-microdissected lung adenocarcinoma cells of early lepidic-types. *Clinical and Translational Medicine, 4*, e24. doi:10.1186/s40169-015-0064-3.

Kawamura, T., Nomura, M., Tojo, H., Fujii, K., Hamasaki, H., Mikami, S., Bando, Y., Kato, H., & Nishimura, T. (2010). Proteomic analysis of laser-microdissected paraffin-embedded tissues: (1) Stage-related protein candidates upon non-metastatic lung adenocarcinoma. *Journal of Proteomics, 73*, 1100–1110.

KEGG PATHWAY Database. (2016). Available from http://www.genome.jp/kegg/pathway.html. 16 Jan 2016.

Koike, T., Yamato, Y., Asamura, H., Tsuchiya, R., Sohara, Y., Eguchi, K., Mori, M., Nakanishi, Y., Goya, T., Koshiishi, Y., Miyaoka, E., & Japanese Joint Committee for Lung Cancer Registration. (2009). Improvements in surgical results for lung cancer from 1989 to 1999 in Japan. *Journal of Thoracic Oncology, 4*, 1364–1369.

LaBaer, J. (2012). Improving international research with clinical specimens: 5 achievable objectives. *Journal of Proteome Research, 11*, 5592–5601.

Leone, A. (2013). Highly sensitive detection of EGFR T790M mutation in pre-TKI specimens of EGFR-mutated NSCLC: In cis, In trans, or a different clone? *Journal of Thoracic Oncology, 8*(3), e26–e27.

Malm, J., Végvári, A., Rezeli, M., Upton, P., Danmyr, P., Nilsson, R., Steinfelder, E., & Marko-Varga, G. (2012). Large scale biobanking of blood – The importance of high density sample processing procedures. *Journal of Proteomics, 76*, 116–124.

Mar, N., & Vredenburgh, J. J. (2015). Dual HER2 blockade in non-small cell lung cancer Harboring a HER2 mutation. *Connecticut Medicine, 79*(9), 531–535.

Marko-Varga, G. (2011). BioBanking – The Holy Grail of novel drug and diagnostic developments. *Journal of Clinical Bioinformatics, 13*, e14.

Marko-Varga, G., Végvári, A., Welinder, C., Rezeli, M., Edula, G., Svensson, K., Belting, M., Laurell, T., & Fehniger, T. E. (2011). Clinical protein science: utilization of biobank resources and examples of current applications. *Journal of Proteome Research, 11*, 5124–5134.

Michalski, A., Cox, J., & Mann, M. (2011). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of Proteome Research, 10*, 1785–1793.

Mitsudomi, T., Morita, S., Yatabe, Y., Negoro, S., Okamoto, I., Tsurutani, J., Seto, T., Satouchi, M., Tada, H., Hirashima, T., Asami, K., Katakami, N., Takada, M., Yoshioka, H., Shibata, K., Kudoh, S., Shimizu, E., Saito, H., Toyooka, S., Nakagawa, K., Fukuoka, M., & Group, W. J. O. (2010). Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): An open label, randomised phase 3 trial. *The Lancet Oncology, 11*(2), 121–128.

Miyauchi, E., Inoue, A., Kobayashi, K., Maemondo, M., Sugawara, S., Oizumi, S., Isobe, H., Gemma, A., Saijo, Y., Yoshizawa, H., Hagiwara, K., Nukiwa, T., & North-East Japan Study Group. (2015). Efficacy of chemotherapy after first-line gefitinib therapy in EGFR mutation-positive advanced non-small cell lung cancer-data from a randomized Phase III study comparing gefitinib with carboplatin plus paclitaxel (NEJ002). *Japanese Journal of Clinical Oncology, 45*(7), 670–676.

Nakamura, H., & Saji, H. (2014). A worldwide trend of increasing primary adenocarcinoma of the lun. *Surgery Today, 44*(6), 1004–1012.

Nakamura, K., Nishio, K., Nishimura, T. (Eds.) & Kato, H (Ed. supervisor). (2012). *Clinical Proteomics*. Tokyo: Kanehara & Co.

Nishimura T., Tojo H. (2014). Mass spectrometry-based protein sequencing platforms. In G. Marko-Varga (Ed.), *Genomics and proteomics for clinical discovery* *and development* (pp. 69–99). (Wang, X (series Ed.) *Translational Bioinformatics 6*). Dordrecht: Springer.

Nomura, M., Fukuda, T., Fujii, K., Kawamura, T., Tojo, H., Kihara, M., Bando, Y., Gazdar, A. F., Tsuboi, M., Oshiro, H., Nagao, T., Ohira, T., Ikeda, N., Gotoh, N., Kato, H., Marko-Varga, G., & Nishimura, T. (2011). Preferential expression of potential markers for cancer stem cells in large cell neuroendocrine carcinoma of the lung. An FFPE proteomic study. *Journal of Clinical Bioinformatics, 1*, e23.

Nowell, P., & Hungerford, D. (1960). A minute chromosome in chronic granulocytic leukemia. *Science, 132*, 1497.

Oxnard, G. R., Nguyen, K. S., & Costa, D. B. (2014). Germline mutations in driver oncogenes and inherited lung cancer risk independent of smoking history. *Journal of the National Cancer Institute, 106*(1), djt361. doi:10.1093/jnci/djt361.

Panchaud, A., Jung, S., Shaffer, S. A., Aitchison, J. D., & Goodlett, D. R. (2011). Faster, quantitative, and accurate precursor acquisition independent from ion count. *Analytical Chemistry, 83*, 2250–2257.

Physikron Mass Spectrometry Systems. (2013). Available from http://www.physikron.com/technology/. 15 Jan 2016.

Prieto, D. A., Hood, B. L., Darfler, M. M., Guiel, T. G., Lucas, D. A., Conrads, T. P., et al. (2005). Liquid Tissue™: Proteomic profiling of formalin-fixed tissues. *BioTechniques, 38*, S32–S35.

RCSB- Research Collaboratory for Structural Bioinformatics. (2015). Protein Data Bank. Available from http://www.rcsb.org/pdb/home/home.do. 16 Jan 2016.

Reactome Pathway Database. (2016). Available from www.reactome.org/. 15 Jan 2016.

Sasaki, H., Shimizu, S., Tani, Y., Maekawa, M., Okuda, K., Yokota, K., Shitara, M., Hikosaka, Y., Moriyama, S., Yano, M., & Fujii, Y. (2012). RET expression and detection of KIF5B/RET gene rearrangements in Japanese lung cancer. *Cancer Medicine, 1*(1), 68–75.

Scagliotti, G. V., Longo, M., & Novello, S. (2009). Nonsmall cell lung cancer in never smokers. *Current Opinion in Oncology, 21*, 99–104.

Semenza, G. (2010). Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics. *Oncogene, 29*, 625–634.

Sequist, L. V., Rolfe, L., & Allen, A. R. (2015). Rociletinib in EGFR-mutated non-small-cell lung cancer. *The New England Journal of Medicine, 373*(6), 578–579.

Shaw, A. T., Camidge, D. R., Engelman, J. A., Solomon, B. J., Kwak, E. L., Clark, J. W., Salgia, R., Shapiro, G., Bang, Y. J., Tan, W., Tye, L., Wilner, K. D., Stephenson, P., Varella-Garcia, M., Bergethon, K., Iafrate, A. J., Ou, S-H. I. (2012). Clinical activity of crizotinib in advanced non-small cell lung cancer (NSCLC) harboring ROS1 gene rearrangement, ASCO Annual Meeting. *Journal of Clinical Oncology*, *30*(Suppl) abstract 7508.

Shtivelman, E., Hensing, T., Simon, G. R., Dennis, P. A., Otterson, G. A., Bueno, R., & Salgia, R. (2014). Molecular pathways and therapeutic targets in lung cancer. *Oncotarget, 5*(6), 1392–1433.

Siegel, R., Ward, E., Brawley, O., & Jemal, A. (2011). Cancer statistics, the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. *CA: A Cancer Journal for Clinicians, 6*, 212–236.

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y., & Mano, H. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature, 448*(7153), 561–566.

STRING-The Search Tool for the Retrieval of Interacting Genes/Proteins database. Available from http://string-db.org/. 15 Jan 2016.

Sun, S., Schiller, J. H., & Gazdar, A. F. (2007). Lung cancer in neversmokers – A different disease. *Nature Reviews Cancer, 7*(10), 778–790.

The data base of Japanese Ministry of Health, Labor and Welfare. (2013) Available from http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai09/kekka3.html. 15 Jan 2016.

The data base of National Cancer Institute at the National Institute of Health. (2013). Available from http://www.cancer.gov/cancertopics/types/lung. 15 Jan 2016.

The National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine, 365*, 395–409.

Thress, K. S., Paweletz, C. P., Felip, E., Cho, B. C., Stetson, D., Dougherty, B., Lai, Z., Markovets, A., Vivancos, A., Kuang, Y., Ercan, D., Matthews, S. E., Cantarini, M., Barrett, J. C., Jänne, P. A., & Oxnard, G. R. (2015). Acquired EGFR C797S mutation mediates resistance to AZD9291 in non-small cell lung cancer harboring EGFR T790M. *Nature Medicine, 21*(6), 560–562.

Toh, C. K., Wong, E. H., Lim, W. T., Leong, S. S., Fong, K. W., Wee, J., & Tan, E. H. (2004). The impact of smoking status on the behavior and survival outcome of patients with advanced non-small cell lung cancer: A retrospective analysis. *Chest, 126*, 1750–1756.

Tomlins, S. A., Laxman, B., Varambally, S., Cao, X., Yu, J., Helgeson, B. E., Cao, Q., Prensner, J. R., Rubin, M. A., Shah, R. B., Mehra, R., & Chinnaiyan, A. M. (2008). Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia, 10*(2), 177–188.

Travis, W. D. (2010). Advances in neuroendocrine lung tumors. *Annals of Oncology, 21*(Suppl. 7), vii 65–vii 71.

Travis, W. D., Linnoila, R. I., Tsokos, M. G., Hitchcock, C. L., Cutler, G. B., Jr., Nieman, L., Chrousos, G., Pass, H., & Doppman, J. (1991). Neuroendocrine tumors of the lung with proposed criteria for large-cell neuroendocrine carcinoma. An ultrastructural, immunohistochemical, and flow cytometric study of 35 cases. *The American Journal of Surgical Pathology, 15*, 529–553.

Travis, W. D., Colby, T. V., Corrin, B., Shimosato, Y., & Brambilla, E. (1999). Introduction. In *Histological typing of lung and pleural tumours* (pp. 1–19). Berlin/ Heidelberg: Springer.

Travis, W. D., Brambilla, E., Müller-Hermelink, H. K., & Harris, C. C. (Eds.). (2004). *Genetics of tumors of lung, pleura, thymus and heart* (pp. 37–38). Lyon: IARC Press.

Travis, W. D., Brambilla, E., Noguchi, M., Nicholson, A. G., Geisinger, K. R., Yatabe, Y., Beer, D. G., Powell, C. A., Riely, G. J., Van Schil, P. E., Garg, K., Austin, J. H., Asamura, H., Rusch, V. W., Hirsch, F. R., Scagliotti, G., Mitsudomi, T., Huber, R. M., Ishikawa, Y., Jett, J., Sanchez-Cespedes, M., Sculier, J. P., Takahashi, T., Tsuboi, M., Vansteenkiste, J., Wistuba, I., Yang, P. C., Aberle, D., Brambilla, C., Flieder, D., Franklin, W., Gazdar, A., Gould, M., Hasleton, P., Henderson, D., Johnson, B., Johnson, D., Kerr, K., Kuriyama, K., Lee, J. S., Miller, V. A., Petersen, I., Roggli, V., Rosell, R., Saijo, N., Thunnissen, E., Tsao, M., & Yankelewitz, D. (2011). International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *Journal of Thoracic Oncology, 6*(2), 244–285.

Varlotto, J. M., Medford-Davis, L. N., Recht, A., Flickinger, J. C., Schaefer, E., Zander, D. S., & DeCamp, M. M. (2011). Should large cell neuroendocrine lung carcinoma be classified and treated as a small cell lung cancer or with other large cell carcinomas? *Journal of Thoracic Oncology, 6*(6), 1050–1058.

Végvári, Á., Rezeli, M., Döme, B., Fehniger, T. E., & Marko-Varga, G. (2011a). Translation science for targeted personalized medicine treatments. In S. Sanders (Ed.), *Selected presentations from the 2011 Sino-American symposium on clinical and translational medicine* (pp. 36–37). Washington, DC: Science/ AAAS.

Végvári, Á., Welinder, C., Lindberg, H., Fehniger, T. E., & Marko-Varga, G. (2011b). Biobank resources for future patient care: Developments, principles and concepts. *Journal of Clinical Bioinformatics, 1*, e24.

Weinstein, I. B. (2001). Addiction to oncogenes—The Achilles heal of cancer. *Science, 297*(5578), 63–64.

Wisniewski, J. R., Ostasiewicz, P., & Mann, M. (2011). High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *Journal of Proteome Research, 10*, 3040–3049.

Wu, X, Fang, X, Zhu, Z, & Wang, X. (2014) Clinical Bioinforatics: A new emerging science of biomarker development. In Marko-Varga, G. (Ed.), *Genomics and proteomics for clinical discovery and development* (pp. 175–191), (Wang, X (series Ed.), *Translational Bioinformatics 6*). Dordrecht: Springer.

Yoshizawa, A., Motoi, N., Riely, G. J., Sima, C. S., Gerald, W. L., Kris, M. G., Park, B. J., Rusch, V. W., & Travis, W. D. (2011). Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases. *Modern Pathology, 24*(5), 653–664.

Yoshizawa, A., Sumiyoshi, S., Sonobe, M., Kobayashi, M., Fujimoto, M., Kawakami, F., Tsuruyama, T., Travis, W. D., Date, H., & Haga, H. (2013). Validation of the IASLC/ATS/ERS lung adenocarcinoma classification for prognosis and association with EGFR and KRAS gene mutations: Analysis of 440 Japanese patients. *Journal of Thoracic Oncology, 8*(1), 52–61.

9 Proteogenomics for the Study of Gastrointestinal Stromal Tumors

Tadashi Kondo

Abstract

Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal tumors of the gastrointestinal tract. Gain-of-function mutations in *KIT* or platelet-derived growth factor receptor alpha (*PDGFRA*) drive most GISTs, and 85 % of GISTs also contain oncogenic mutations in one of two receptor tyrosine kinases. The advent of tyrosine kinase inhibitors has had a significant impact on the clinical practices for GISTs. However, tumors in more than 80 % of GIST patients acquire resistance against treatments with tyrosine kinase inhibitors; thus, driver mechanisms of secondary resistance as well as biomarkers for early detection of recurrence have been explored for better clinical outcomes. Proteomics is a versatile and straightforward approach to finding the molecular basis of malignancies as well as the innovative seeds for clinical applications. Comprehensive genome, epigenome, and transcriptome data have already been obtained and examined together in GISTs, and proteome data has a unique additional value in multi-omics studies. Various types of samples were examined using proteomics modalities in GIST, suggest the promising utility of proteomic approaches.

Keywords

Gastrointestinal stromal tumor • GIST • Biomarker • Molecular targeting drug

9.1 Introduction

Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal tumors of the gastrointestinal tract. The incidence of GIST is 6.8 per million, and 3300–6000 new GIST cases are reported per year in the United States (Corless

T. Kondo (✉)
Division of Rare Cancer Research, National Cancer Center Research Institute,
5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
e-mail: takondo@ncc.go.jo

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_9

and Heinrich 2008; Tran et al. 2005). GISTs can occur anywhere along the gastrointestinal tract, but they predominantly arise in the stomach (60–70 %) and small intestine (20–30 %) (Miettinen and Lasota 2001). The original site of the tumor correlates with the clinical features; approximately 20–25 % of gastric GISTs and 40–50 % of small intestinal GISTs are clinically aggressive (Joensuu 2006, 2008; Miettinen and Lasota 2006; Emory et al. 1999). The most common presentations of these tumors include bleeding from the gastrointestinal tract, acute abdomen pain due to tumor rupture, appendicitis-like pain, and obstruction as well as fatigue, dysphagia, and satiety. GIST cells originate from the interstitial cells of Cajal, which are pacemaker cells that regulate peristalsis in the digestive tract (Hirota et al. 1998; Kindblom et al. 1998). Gain-of-function mutations in *KIT* or platelet-derived growth factor receptor alpha (*PDGFRA*) drive most GISTs, and 85 % of GISTs also contain oncogenic mutations in one of two receptor tyrosine kinases (Corless and Heinrich 2008; Miettinen and Lasota 2006). Mutations in these tyrosine kinases play key roles in GIST pathogenesis and proliferation. Therefore, molecular targeted therapy with KIT/PDGFRA tyrosine kinase inhibitors such as imatinib mesylate (Gleevec, Novartis Pharmaceuticals) has a beneficial effect in a significant portion of GIST patients (Corless et al. 2011; Dematteo et al. 2002; Verweij et al. 2004). Because GISTs are highly resistant to conventional chemotherapy (Rubin et al. 2007; Patel et al. 1998, 2001; Trent et al. 2003), the advent of tyrosine kinase inhibitors has had a significant impact on the clinical practices for these tumors. However, tumors in more than 80 % of GIST patients acquire resistance against treatments with tyrosine kinase inhibitors; thus, driver mechanisms of secondary resistance as well as biomarkers for early detection of recurrence have been explored for better clinical outcomes. The prognosis of GIST after surgery is generally favorable (Joensuu et al. 2012), and prognostic modalities will contribute to the optimized indication for adjuvant treatments using tyrosine kinase inhibitors in GISTs. Calculating the risk of progression for a newly diagnosed primary GIST currently relies on mitotic index, tumor size, and tumor location (Corless and Heinrich 2008). Looking forward, novel prognostic biomarkers have potential to improve risk stratification therapy.

9.2 Proteomic Approach to GIST

Proteomics is a versatile and straightforward approach to finding the molecular basis of malignancies, because the proteome is a functional translation of the genome that regulates the behavior of cells. Proteomics provides considerable information about the overall features of proteins that cannot be obtained otherwise, such as protein expression levels, protein–protein interactions, post-translational modifications, and protein localization. Although information about these protein features should be encoded in the genome, they are not predictable from genome data yet, and we have to examine the proteome itself to obtain the overall features of expressed proteins. Proteomics modalities have been employed in cancer research, and intriguing proteins have been identified as biomarker candidates as well as possible therapeutic targets in sarcomas (Kondo et al. 2012). In GIST studies looking toward identification of novel innovative seeds for clinical applications, various types of samples were examined using proteomics modalities. Materials for proteomic studies have included surgically resected tissues, GIST cell lines, and conditioned medium of cultured GIST cells.

9.3 Proteomics for Surgically Resected Tissues

Biomarkers to predict post-operative recurrence have the potential to improve the guidelines for adjuvant therapy using tyrosine kinase inhibitors. Aiming to develop biomarkers for post-surgical recurrence, Suehara et al. examined protein expression in primary tumor tissues of GIST (Suehara et al. 2008). Proteins in primary tumor tissues were compared between patients with different recurrence statuses after operation: the patients who had recurrence within 1 year after surgery, and the patients with no recurrence more than 2 years after surgery (Suehara et al. 2008). To create protein expression profiles, two-dimensional difference in gel electrophoresis (2D-DIGE) was employed. In 2D-DIGE, protein samples are labeled with fluorescent dyes, mixed, and separated on the same 2D-PAGE gel (Unlu et al. 1997; Shaw et al. 2003; Kondo and Hirohashi 2007). After electrophoresis, images are obtained by laser scanning, and multiple images for individual samples are compared. By comparative studies using 2D-DIGE, 1513 protein spots were observed and 25 unique proteins were identified that exhibited significant differences in expression level between the patient groups. Among these differentially expressed proteins, one protein, pfetin (potassium channel tetramerization domain containing 12, KCTD12), was chosen for further examination. Pfetin was originally discovered as a protein with unique expression in the fetal cochlea (Resendes et al. 2004). However, in this GIST proteomic study, pfetin was detected in multiple protein spots, the signal intensity of which was higher in the GIST patients who had no evidence of metastasis more than 2 years after surgery (Suehara et al. 2008). The association of pfetin expression with favorable prognosis was confirmed in 210 GIST cases by immunohistochemistry (Suehara et al. 2008). The prognostic value of pfetin was independently significant among clinical and pathological parameters, and pfetin exhibited prognostic value even in samples grouped by risk classification. Moreover, the prognostic utility of pfetin was further validated in additional GIST cases by immunohistochemistry (Kondo et al. 2013; Hasegawa et al. 2013; Orita et al. 2014; Kikuta et al. 2010; Kubota et al. 2011, 2012). These observations suggested that pfetin may be a useful marker for predicting the likelihood of recurrence in GIST patients and for identifying which patients should be recommended to avoid adjuvant treatment.

Biomarkers for recurrence after surgery were also discovered by 2D-DIGE with a large format gel electrophoresis in another study (Kondo and Hirohashi 2007). Kikuta et al. examined primary tumor tissues of GIST patients who had different prognoses after surgical operation (Kikuta et al. 2012). They observed 3260 protein spots, and identified 25 unique proteins exhibiting differential expression between the patient groups. Among them, they focused on the expression of ATP-dependent RNA helicase DDX39 (DDX39). DDX39 belongs to the DEAD (DExD/H) box RNA helicases which unwind double-stranded RNA (Linder et al. 1989), and was found to be overexpressed in lung cancer (Sugiura et al. 2007). The expression of DDX39 was upregulated in GIST patients who had metastasis within 1 year after surgical operation, compared with patients with no metastasis more than 2 years after surgery (Kikuta et al. 2012). GIST patients with higher level of DDX39 expression had lower probability of disease-free survival; the prognostic value of DDX39 was confirmed in 72 independent GIST cases by immunohistochemistry (Kubota et al. 2012). These results suggested the possible use of DDX39 as a novel biomarker for risk stratification therapy.

In GISTs, clinical outcomes are significantly different depending on the primary tumor site; GISTs of the small intestine exhibit more aggressive behavior than those of the stomach, despite similar size and mitotic activity (Joensuu 2006, 2008; Miettinen and Lasota 2006; Emory et al. 1999). Therefore, the location of the primary site is one of the factors considered in risk stratification schemes. Suehara et al. examined the differential protein expression between GIST originating in the stomach and GIST originating in the small intestine (Suehara et al. 2009). Using 2D-DIGE, they observed 1411 protein spots, and identified 72 unique proteins with differential expression according to original tumor site. They examined global expression of mRNA, comparing proteomic and transcriptomic data. However, as 2D-DIGE generated multiple protein spots from single genes, the results of association studies between the proteome and transcriptome did not yield conclusive results, and pairs of proteins and mRNAs with obvious concordance were not identified. Ichikawa et al. also compared primary GIST tumor tissues from the stomach to those from the small intestine by both proteomic and transcriptomic approaches (Ichikawa et al. 2014). The proteins extracted from primary tumor tissues were first separated by SDS-PAGE according to their molecular weight. After gel electrophoresis, separated proteins were recovered as tryptic digests and subjected to mass spectrometry. A total of 2555 unique proteins were observed, and the mRNA data corresponding to those proteins were extracted from the Gene Expression Omnibus. Protein and mRNA were compared between the stomach GISTs and small intestine GISTs, and 18 unique proteins were identified as being commonly differentially expressed. Among the 18 proteins identified, promyelocytic leukemia protein (PML) was chosen for further examination. PML was originally identified as a fusion partner of retinoic acid receptor-alpha in acute promyelocytic leukemia (Melnick and Licht 1999), functioning as a tumor suppressor (Salomoni et al. 2008). Loss of PML was reported in various malignancies including breast cancer, gastric cancer, and small cell lung cancer (Gurrieri et al. 2004). After proteomic study, the prognostic value of PML was confirmed in 254 GIST cases by immunohistochemistry (Ichikawa et al. 2014). The prognostic value of PML was significant in stomach GIST, and PML exhibited prognostic value even in the samples grouped by risk classification (Ichikawa et al. 2014). These results demonstrate the possible use of PML as a novel biomarker for risk stratification therapy. Further validation studies using additional GIST cases will be required for developing clinical applications of PML in GIST, and the functional relevance of unique PML expression should also be explored.

The evaluation of response to imatinib treatment is critical to establish a therapeutic strategy for GIST. However, size-based response criteria such as the World Health Organization criteria or the Response Evaluation Criteria in Solid Tumors may underestimate the response (Scaife et al. 2003). Moreover, the histological/pathological response of GIST to imatinib therapy is variable, heterogeneous, and does not correlate well with clinical response. To explore the molecular effects of imatinib on responding GISTs, Luca et al. examined the proteomic features of GIST tissues resected after imatinib treatment. They separated proteins by SDS-PAGE and an Agilent 3100 OFFGEL fractionator (Agilent Technologies, Santa Clara, CA) according to molecular weight and isoelectric point, respectively. The separated proteins were digested to tryptic peptides and subjected to mass spectrometry. An elevated amount of stem cell growth factor (SCGF), a hematopoietic growth factor with a role in the development of erythroid and myeloid progenitors, was detected in the imatinib-treated GIST cells. SCGF localized in the stromal component area of GIST tissues, likely due to the imatinib-induced inflammation response. These results suggest an important functional role for SCGF in the response area, and the possible utility of SCGF as a biomarker to predict the effects of imatinib treatment. Further validation and functional studies will be required for developing clinical applications of SCGF in GISTs.

GISTs are typically diagnosed in adults over the age of 40, with a peak incidence between 60 and 70 years of age (DeMatteo et al. 2000). Pediatric GISTs are extremely rare, accounting for 1–2 % of all GIST cases, and annual incidence of GIST cases is 0.02 per million in children under the age of 14 (Stiller 2007). Pediatric GISTs have unique characteristics; they occur preferentially in females as multiple nodules, having either an epithelioid or a mixed spindle and epithelioid morphology (Miettinen et al. 2005; Prakash et al. 2005). In pediatric GISTs, risk of metastasis is low, the tumors often lack mutations in *KIT* and *PDGFRA*, and the efficacy of kinase inhibitors in pediatric GISTs has not been well defined. To explore the molecular background of pediatric GISTs, Agaram et al. compared global mRNA expression of primary tumor tissues in adult and pediatric GIST (Agaram et al. 2008). They reported 14 genes were differently expressed between pediatric and adult GIST patients, and identified a gene expression signature of pediatric GIST. In addition to transcriptomic studies, they also examined phosphorylation status of receptor tyrosine kinases using an antibody array (Human phosphor-RTK array kit, R&D Systems, Inc., Minneapolis, MN)

(Agaram et al. 2008). They demonstrated that adult GISTs exhibited phosphorylation of PDGFRB, EGFR, and FGFR2a without KIT activation, but pediatric GISTs had phosphorylated KIT and a weakly phosphorylated EGFR. These observations may explain the unique clinical features of pediatric GIST. Other possible effects of the identified proteins on the specific clinical features of pediatric GIST are worth further investigating.

Mutations in the KIT gene may account for unique molecular characteristics of GISTs and the effects of these mutations on the proteome are of interest. Choi et al. compared primary tumor tissues of GIST patients with or without KIT mutations at exon 11 (Choi et al. 2003). Using 2D-PAGE, more than 1000 protein spots were observed, and increased expression of High Mobility Group Box 1 (HMG1) was observed in the GISTs with KIT mutations compared with the GISTs without KIT mutations. HMG1 was originally identified as a chromosomal DNA-binding protein (Bustin 1999). Because HMG1 supports transcription of genes interacting with transcription factors, its overexpression may influence transcriptional activity in GIST cells with KIT mutation. The overexpression of HMG1 was reported in various types of malignancies (Tang et al. 2010; Kang et al. 2013). HMG1 was considered as a target for cancer therapy (Lotze and DeMarco 2003), and as a biomarker candidate for poor prognosis (Shi et al. 2015; Ladoire et al. 2015) and response to treatment (Shrivastava et al. 2015). These observations suggest the presence of shared mechanisms underlying common features of the different types of malignancies, and the possible application of HMG1 for novel therapeutic strategies in GISTs. The mechanisms by which KIT mutations upregulate HMG1 expression are worth investigating in further studies.

GISTs are characterized by gain-of-function mutations in *KIT* or *PDGFRA*, and 85 % of GISTs contain oncogenic mutations in one of these two receptor tyrosine kinases (Corless and Heinrich 2008; Miettinen and Lasota 2006). Mutations of these two genes are mutually exclusive, and these mutations are regarded as alternative oncogenic mechanisms in GISTs. Small subsets of GISTs have no mutation in either of these two genes; those GISTs have mutations in other genes. GISTs with different gene mutations likely have different proteomic signatures, which should include distinct biomarkers and target candidates. To explore the proteomic features of GISTs with different gene mutations, Kang et al. examined the protein expression profiles of primary tumor tissues with different mutation types (Kang et al. 2006). Moreover, they also examined the proteins with known expression patterns associated with risk classification. To create protein expression profiles, they employed 2D-PAGE. The comparative study between GIST with *KIT* mutations, GIST with *PDGFRA* mutations, and GIST lacking either mutation resulted in the identification of proteins whose expression levels were unique to each group of GIST with different mutation types. The overexpression of septin and HSP27 were unique to GIST with KIT mutations, expression of keratin 10 was unique to GIST with *PDGFRA* mutations, and expression of annexin V was unique to GIST lacking either mutation. The proteins with higher expression in GIST with high risk for recurrence included annexin V, HMGB1, C13orf2, glutamate dehydrogenase 1, and fibrinogen beta chain. C13orf2 is an alias of pfetin (KCTD12). In the study by Suehara et al., higher expression of pfetin was associated with favorable prognosis (Fowler et al. 2013), but these results are discordant to the study by Kang et al. (2006). The cause-and-effect relationship between kinase mutations and differential protein expression has not yet been examined, and the molecular mechanisms underlying differential regulation of the identified proteins are worth investigating in future studies.

As previously mentioned, GISTs are mesenchymal tumors with unique genetic characteristics such as frequent mutations in *KIT* and *PDGFRA*. Identification of proteins unique to GIST cells will yield further insight into the effects of specific mutations on cell behavior. Suehara et al. created the protein expression profiles of primary tumor tissues of various types of sarcomas including GIST (Suehara et al. 2006). Using 2D-DIGE, they observed more than 1200 protein spots across 80 sarcoma tissue samples, and identified proteins unique to the histological subtypes. They reported 10 proteins by which the cross-validation error rate for GIST with the other sarcomas was minimal. The specificity and diagnostic utility of the identified proteins are worth examining in the additional samples.

Although GISTs have similar morphology to leiomyosarcomas, these two sarcomas are clinically distinct (Clary et al. 2001; Fletcher et al. 2002). In advanced GIST patients, chemotherapeutic agents result in response rate of only 0–10 % (Rubin et al. 2007; Patel et al. 1998, 2001; Trent et al. 2003), and the response rate of imatinib treatments was greater than 50 % (Corless et al. 2011; Dematteo et al. 2002; Verweij et al. 2004). In contrast, in advanced leiomyosarcoma patients, the response rate of combination therapy with chemotherapeutic agents gemcitabine and docetaxel was 53 % (Hensley et al. 2002), and imatinib treatment was not beneficial (Silvestris et al. 2005). Thus, differential diagnosis is critical for the treatment of GISTs and leiomyosarcomas. Although transcriptomic studies reported unique gene expression patterns in GISTs compared with leiomyosarcomas (Yang et al. 2010), there was no investigation of this issue using a proteomic approach. Yang et al. first examined global protein expression between GISTs and leiomyosarcomas using a reverse-phase protein lysate array. Unique expression of E-cadherin was identified in leiomyosarcoma (Yang et al. 2010). Moreover, the transcription factor Slug was reported as a possible regulatory gene by transcriptomic experiments and in vitro function studies. In malignant tumors of epithelial origin, the suppression of E-cadherin is associated with the epithelial to mesenchymal transition (EMT), which accounts for increased invasion and metastasis during tumor progression. In the reverse process of EMT, mesenchymal to epithelial reverting transition (MErT), E-cadherin plays an important role, and certain kinds of soft-tissue sarcomas with epithelioid features expressed E-cadherin (Sato et al. 1999). Yang's report suggested the utility of E-cadherin as a differential diagnosis biomarker between GISTs and leiomyosarcoma and the utility of Slug as a potential therapeutic target in leiomyosarcoma. It is worth pursuing these proteins further for developing clinical applications of E-cadherin and Slug.

9.4 Proteomics for Cultured Cells

Imatinib treatment yields considerable benefits for GIST patients; adjuvant imatinib improves both recurrence-free and overall survival (Dematteo et al. 2009; Demetri et al. 2002). However, secondary resistance occurs in more than 80 % of patients after treatment with imatinib. Although secondary mutations in *KIT* or *PDGFRA* that interfere with drug binding were observed (Chen et al. 2004; Wardelmann et al. 2005; Antonescu et al. 2005; Heinrich et al. 2006), they are highly heterogeneous, even within different areas of the same tumor, and the clinical benefits of second-and third-line drugs (*e.g.*, sunitinib, regorafenib) are quite limited (Demetri et al. 2013). To explore the molecular backgrounds of resistance against imatinib treatment, Takahashi et al. examined global expression of tyrosine-phosphorylated proteins in GIST-T1 cells (Takahashi et al. 2013). GIST-T1 cells have a 57-nucleotide in-frame deletion in KIT exon 11, and are used for in vitro studies of GIST (Taguchi et al. 2002). Protein was extracted from the GIST-T1 cells, and tyrosine-phosphorylated peptides were purified using a specific antibody. Using the isobaric tags for relative and absolute quantitation (iTRAQ) method, 171 tyrosine phosphorylation sites spanning 134 proteins were observed, and 26 tyrosine-phosphorylated proteins whose expression levels were altered by imatinib treatment were identified. Among these, they focused on tyrosine-protein kinase FYN (FYN) and focal adhesion kinase 1 (FAK) for further investigation. FYN plays an important role in the signaling pathway of integrin and PI3K (Timokhina et al. 1998; Linnekin et al. 1997), and FAK transduces signals from integrin and growth factor receptors to downstream pathways (Serrels et al. 2012; Cox et al. 2006). They demonstrated that activation of these two kinases contributes to resistance to imatinib treatment, and that inhibition of these kinases resensitizes GIST cells to imatinib treatment (Takahashi et al. 2013). The constitutive phosphorylation of FAK was observed in the imatinib-resistant GIST-T1 cells, and a FAK-specific TAG372 inhibitor decreased the viability of these GIST cells with and without imatinib treatment. These observations suggest the novel utility of signal transduction pathways including FYN and FAK as potential targets for GIST therapy.

To explore the cellular changes associated with imatinib treatment and secondary resistance, Nagata et al. examined the phosphorylation of proteins in GIST882 cells (GIST882), GIST882 cells under treatment with imatinib (GIST882-IM), and GIST882 cells with secondary imatinib resistance (GIST882-R) (Nagata et al. 2015). Proteins extracted from these GIST882 cells were digested with trypsin, and phosphorylated peptides were purified by titania-based affinity chromatography or immunoprecipitation by anti-phosphotyrosine antibody. The purified phosphopetides were subjected to mass spectrometric analysis. They observed 1036 peptides containing phosphorylated Ser, Thr, and/or Tyr residues enriched by the titania-based method, and 210 phosphotyrosine-containing peptides by the immunoprecipitation method. These studies found that resistance to imatinib might result from activation of alternative receptor type kinases, including EGFR, and their downstream signaling pathways. Expression of KIT and EGFR was upregulated in GIST882-R cells compared with GIST882 and GIST882-IM, and treatment with the EGFR inhibitor gefitinib had anti-proliferative effects on GIST882-R. Although the overexpression of EGFR was observed in the most GIST cases examined by immunohistochemistry (Lopes and Bacchi 2007), the correlation between the expression level of EGFR and the survival benefits was not confirmed in GIST (Jiang et al. 2012). Thus, the clinical significance of this in vitro study should be further explored using clinical samples.

9.5 Proteomics for Conditioned Medium of Cultured Cells

To investigate the secreted, shed, or leaked proteins from GIST cells, Berglund et al. investigated proteins in the conditioned medium of GIST882. GIST882 is the first established immortalized GIST cell line, harboring a KIT mutation and imatinib sensitivity. GIST882 possesses a homozygous missense mutation in exon 13 of the c-kit gene, and are commonly used for in vitro studies of GIST (Tuveson et al. 2001). Cell lines are useful resources for identifying biomarkers in body fluid, because proteins with low expression levels can be concentrated and examined by proteomic modalities. Moreover, the proteins derived exclusively from tumor cells can be recovered and subjected to proteomic studies. Berglund et al. observed the release of 764 proteins from GIST882. They found that release of nuclease-sensitive element-binding protein 1 (Y-box binding protein 1) was induced by treatment with imatinib. As Y-box binding protein 1 is involved in the acquisition of global drug resistance through increased MDR1 expression (Basaki et al. 2007), these observations may yield clues to understanding the molecular mechanisms underlying acquired resistance in GIST patients.

9.6 Perspectives of Proteomic Approach to GIST

Various proteomics modalities have already been applied to the study of GIST, and many intriguing proteins were reported in the aforementioned studies. Although the biological and clinical significance of the identified proteins should be further functionally verified and validated in independent GIST cases, these results suggest the promising utility of proteomic approaches to the study of GIST. Proteomic modalities applied to GIST study included 2D-DIGE, 2D-PAGE, mass spectrometry with or without iTRAQ method, antibody arrays, and reverse-phase protein arrays. As the observable proteomic features largely depend on the proteomics methods employed, we can expect novel findings using proteomics modalities that have not yet been utilized. The future integration of proteome data with the other multi-omics data is also expected. Comprehensive genome, epigenome, and transcriptome data have been already obtained and examined together in GISTs (Haller et al. 2015; Okamoto et al. 2012; Yamaguchi et al. 2008; Brenca et al. 2015; Saponara et al. 2015; Hara et al. 2015; Arne et al. 2011), and proteome data has unique value in multi-omics studies.

In GIST proteomics, many important research topics have not yet been investigated. Those include the development of plasma biomarkers for disease monitoring, the elucidation of resistance mechanisms for tyrosine kinase inhibitors other than imatinib, and the comprehensive understanding of a variety of mutations in tyrosine kinase genes. To conduct the research toward clinical applications efficiently and effectively, collaboration between clinical, academic, and industry groups is necessary from early stages of research. This is a general concern in clinical-problem oriented research.

The number of tissue samples examined in proteomics studies has generally been low. The frozen tissue samples required for conventional proteomics are not routinely stored in hospitals. In addition, because of the low prevalence of GIST patients, especially when the GIST cases are stratified, it takes a long time to collect samples in a prospective way. One possible solution for these issues may be the use of a biobanking system. Biobanking could eventually contribute to research of rare cancers such as GIST. Moreover, since hundreds of GIST cases were successfully examined for validation studies using immunohistochemistry (Kondo et al. 2013; Hasegawa et al. 2013; Orita et al. 2014; Kikuta et al. 2010; Kubota et al. 2011, 2012), the use of formalin-fixed, paraffin-embedded (FFPE) samples for proteomics should be seriously considered for GIST research (Fowler et al. 2013). Global expression data obtained from a small number of samples does not generate conclusive results, and integration of meta-data should be considered. Therefore, standardizing proteomic methods and data is important, especially for the study of rare cancers such as GIST (Deutsch et al. 2015). As for proteomic study of tumor tissues, the use of laser microdissection should be considered, because of the inter- and intra-tumor heterogeneity of resistance mutations and gene amplification in GIST tissues (Liegl et al. 2008).

Mechanisms of acquired resistance were explored using cultured cells, and several intriguing proteins and signal transduction pathways were identified (Takahashi et al. 2013; Nagata et al. 2015). The use of biopsied samples in the recurrent tumor tissues is worth considering for validation of the results obtained by in vitro studies. Cultured cells are useful resources to examine the effects of cancer drugs on tumor cells and to investigate the molecular basis of resistance. However, only two GIST cell lines were examined in the proteomic studies for GIST (Takahashi et al. 2013; Nagata et al. 2015). These GIST cell lines are not deposited in public cell banks, and should be more accessible for the researchers. Moreover, the establishment of additional GIST cell lines is necessary to accurately represent the complexity of GIST disease backgrounds. As the interaction and distribution of tumor cells and stromal cells in the tumor microenvironment affect a range of cellular functions (Yamada and Cukierman 2007; Wang et al. 2002; Vaira et al. 2010; Ridky et al. 2010), the use of three-dimensional organotypic culturing systems should be considered in GIST studies.

Acknowledgments This work was supported by supported by the Practical Research for Innovative Cancer Control (15ck0106089h0002) from Japan Agency for Medical Research and Development (AMED), and the National Cancer Center Research and Development Fund (26-A-9).

References

Agaram, N. P., Laquaglia, M. P., Ustun, B., Guo, T., Wong, G. C., Socci, N. D., Maki, R. G., DeMatteo, R. P., Besmer, P., & Antonescu, C. R. (2008). Molecular characterization of pediatric gastrointestinal stromal tumors. *Clinical Cancer Research, 14*(10), 3204–3215. doi:10.1158/1078-0432.CCR-07-1984.

Antonescu, C. R., Besmer, P., Guo, T., Arkun, K., Hom, G., Koryotowski, B., Leversha, M. A., Jeffrey, P. D., Desantis, D., Singer, S., Brennan, M. F., Maki, R. G., & DeMatteo, R. P. (2005). Acquired resistance to imatinib in gastrointestinal stromal tumor occurs through secondary gene mutation. *Clinical Cancer Research, 11*(11), 4182–4190. doi:10.1158/1078-0432.CCR-04-2245.

Arne, G., Kristiansson, E., Nerman, O., Kindblom, L. G., Ahlman, H., Nilsson, B., & Nilsson, O. (2011). Expression profiling of GIST: CD133 is associated with KIT exon 11 mutations, gastric location and poor prognosis. *International Journal of Cancer, 129*(5), 1149–1161. doi:10.1002/ijc.25755.

Basaki, Y., Hosoi, F., Oda, Y., Fotovati, A., Maruyama, Y., Oie, S., Ono, M., Izumi, H., Kohno, K., Sakai, K., Shimoyama, T., Nishio, K., & Kuwano, M. (2007). Akt-dependent nuclear localization of Y-box-binding protein 1 in acquisition of malignant characteristics by human ovarian cancer cells. *Oncogene, 26*(19), 2736–2746. doi:10.1038/sj.onc.1210084.

Brenca, M., Rossi, S., Polano, M., Gasparotto, D., Zanatta, L., Racanelli, D., Valori, L., Lamon, S., Dei Tos, A. P., & Maestro, R. (2015). Transcriptome sequencing identifies ETV6-NTRK3 as a gene fusion involved in GIST. *The Journal of Pathology*. doi:10.1002/path.4677.

Bustin, M. (1999). Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. *Molecular and Cellular Biology, 19*(8), 5237–5246.

Chen, L. L., Trent, J. C., Wu, E. F., Fuller, G. N., Ramdas, L., Zhang, W., Raymond, A. K., Prieto, V. G., Oyedeji, C. O., Hunt, K. K., Pollock, R. E., Feig, B. W., Hayes, K. J., Choi, H., Macapinlac, H. A., Hittelman, W., Velasco, M. A., Patel, S., Burgess, M. A., Benjamin, R. S., & Frazier, M. L. (2004). A missense mutation in KIT kinase domain 1 correlates with imatinib resistance in gastrointestinal stromal tumors. *Cancer Research, 64*(17), 5913–5919. doi:10.1158/0008-5472.CAN-04-0085.

Choi, Y. R., Kim, H., Kang, H. J., Kim, N. G., Kim, J. J., Park, K. S., Paik, Y. K., Kim, H. O., & Kim, H. (2003). Overexpression of high mobility group box 1 in gastrointestinal stromal tumors with KIT mutation. *Cancer Research, 63*(9), 2188–2193.

Clary, B. M., DeMatteo, R. P., Lewis, J. J., Leung, D., & Brennan, M. F. (2001). Gastrointestinal stromal tumors and leiomyosarcoma of the abdomen and retroperitoneum: A clinical comparison. *Annals of Surgical Oncology, 8*(4), 290–299.

Corless, C. L., & Heinrich, M. C. (2008). Molecular pathobiology of gastrointestinal stromal sarcomas. *Annual Review of Pathology, 3*, 557–586. doi:10.1146/annurev.pathmechdis.3.121806.151538.

Corless, C. L., Barnett, C. M., & Heinrich, M. C. (2011). Gastrointestinal stromal tumours: Origin and molecular oncology. *Nature Reviews Cancer, 11*(12), 865–878.

Cox, B. D., Natarajan, M., Stettner, M. R., & Gladson, C. L. (2006). New concepts regarding focal adhesion kinase promotion of cell migration and proliferation. *Journal of Cellular Biochemistry, 99*(1), 35–52. doi:10.1002/jcb.20956.

DeMatteo, R. P., Lewis, J. J., Leung, D., Mudan, S. S., Woodruff, J. M., & Brennan, M. F. (2000). Two hundred gastrointestinal stromal tumors: Recurrence patterns and prognostic factors for survival. *Annals of Surgery, 231*(1), 51–58.

Dematteo, R. P., Heinrich, M. C., El-Rifai, W. M., & Demetri, G. (2002). Clinical management of gastrointestinal stromal tumors: Before and after STI-571. *Human Pathology, 33*(5), 466–477. doi:S0046817702000163 [pii].

Dematteo, R. P., Ballman, K. V., Antonescu, C. R., Maki, R. G., Pisters, P. W., Demetri, G. D., Blackstein, M. E., Blanke, C. D., von Mehren, M., Brennan, M. F., Patel, S., McCarter, M. D., Polikoff, J. A., Tan, B. R., Owzar, K., & American College of Surgeons Oncology Group Intergroup Adjuvant GST. (2009). Adjuvant imatinib mesylate after resection of localised, primary gastrointestinal stromal tumour: A randomised, double-blind, placebo-controlled trial. *Lancet, 373*(9669), 1097–1104. doi:10.1016/S0140-6736(09)60500-6.

Demetri, G. D., von Mehren, M., Blanke, C. D., Van den Abbeele, A. D., Eisenberg, B., Roberts, P. J., Heinrich, M. C., Tuveson, D. A., Singer, S., Janicek, M., Fletcher, J. A., Silverman, S. G., Silberman, S. L., Capdeville, R., Kiese, B., Peng, B., Dimitrijevic, S., Druker, B. J., Corless, C., Fletcher, C. D., & Joensuu, H. (2002). Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *The New England Journal of Medicine, 347*(7), 472–480. doi:10.1056/NEJMoa020461347/7/472 [pii].

Demetri, G. D., Reichardt, P., Kang, Y. K., Blay, J. Y., Rutkowski, P., Gelderblom, H., Hohenberger, P., Leahy, M., von Mehren, M., Joensuu, H., Badalamenti, G., Blackstein, M., Le Cesne, A., Schoffski, P., Maki, R. G., Bauer, S., Nguyen, B. B., Xu, J., Nishida, T., Chung, J., Kappeler, C., Kuss, I., Laurent, D., Casali, P. G., & investigators Gs. (2013). Efficacy and safety of regorafenib for advanced gastrointestinal stromal tumours after failure of imatinib and sunitinib (GRID): An international, multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet, 381*(9863), 295–302. doi:10.1016/S0140-6736(12)61857-1.

Deutsch, E. W., Albar, J. P., Binz, P. A., Eisenacher, M., Jones, A. R., Mayer, G., Omenn, G. S., Orchard, S., Vizcaino, J. A., & Hermjakob, H. (2015). Development of data representation standards by the human proteome organization proteomics standards initiative. *Journal of the American Medical Informatics Association: JAMIA, 22*(3), 495–506. doi:10.1093/jamia/ocv001.

Emory, T. S., Sobin, L. H., Lukes, L., Lee, D. H., & O'Leary, T. J. (1999). Prognosis of gastrointestinal smooth-muscle (stromal) tumors: Dependence on anatomic site. *The American Journal of Surgical Pathology, 23*(1), 82–87.

Fletcher, C. D., Berman, J. J., Corless, C., Gorstein, F., Lasota, J., Longley, B. J., Miettinen, M., O'Leary, T. J., Remotti, H., Rubin, B. P., Shmookler, B., Sobin, L. H., & Weiss, S. W. (2002). Diagnosis of gastrointestinal stromal tumors: A consensus approach. *Human Pathology, 33*(5), 459–465. doi:S0046817702000151 [pii].

Fowler, C. B., O'Leary, T. J., & Mason, J. T. (2013). Toward improving the proteomic analysis of formalin-fixed, paraffin-embedded tissue. *Expert Review of Proteomics, 10*(4), 389–400. doi:10.1586/14789450.2013.820531.

Gurrieri, C., Capodieci, P., Bernardi, R., Scaglioni, P. P., Nafa, K., Rush, L. J., Verbel, D. A., Cordon-Cardo, C., & Pandolfi, P. P. (2004). Loss of the tumor suppressor PML in human cancers of multiple histologic origins. *Journal of the National Cancer Institute, 96*(4), 269–279.

Haller, F., Zhang, J. D., Moskalev, E. A., Braun, A., Otto, C., Geddert, H., Riazalhosseini, Y., Ward, A., Balwierz, A., Schaefer, I. M., Cameron, S., Ghadimi, B. M., Agaimy, A., Fletcher, J. A., Hoheisel, J., Hartmann, A., Werner, M., Wiemann, S., & Sahin, O. (2015). Combined DNA methylation and gene expression profiling in gastrointestinal stromal tumors reveals hypomethylation of SPP1 as an independent prognostic factor. *International Journal of Cancer, 136*(5), 1013–1023. doi:10.1002/ijc.29088.

Hara, R., Kikuchi, H., Setoguchi, T., Miyazaki, S., Yamamoto, M., Hiramatsu, Y., Kamiya, K., Ohta, M., Baba, S., & Konno, H. (2015). Microarray analysis reveals distinct gene set profiles for gastric and intestinal gastrointestinal stromal tumors. *Anticancer Research, 35*(6), 3289–3298.

Hasegawa, T., Asanuma, H., Ogino, J., Hirohashi, Y., Shinomura, Y., Iwaki, H., Kikuchi, H., & Kondo, T. (2013). Use of potassium channel tetramerization domain-containing 12 as a biomarker for diagnosis and prognosis of gastrointestinal stromal tumor. *Human Pathology, 44*(7), 1271–1277. doi:10.1016/j.humpath.2012.10.013. S0046-8177(12)00381-4 [pii].

Heinrich, M. C., Corless, C. L., Blanke, C. D., Demetri, G. D., Joensuu, H., Roberts, P. J., Eisenberg, B. L., von Mehren, M., Fletcher, C. D., Sandau, K., McDougall, K., Ou, W. B., Chen, C. J., & Fletcher, J. A. (2006). Molecular correlates of imatinib resistance in gastrointestinal stromal tumors. *Journal of Clinical Oncology, 24*(29), 4764–4774. doi:10.1200/JCO.2006.06.2265.

Hensley, M. L., Maki, R., Venkatraman, E., Geller, G., Lovegren, M., Aghajanian, C., Sabbatini, P., Tong, W., Barakat, R., & Spriggs, D. R. (2002). Gemcitabine and docetaxel in patients with unresectable leiomyosarcoma: Results of a phase II trial. *Journal of Clinical Oncology, 20*(12), 2824–2831.

Hirota, S., Isozaki, K., Moriyama, Y., Hashimoto, K., Nishida, T., Ishiguro, S., Kawano, K., Hanada, M., Kurata, A., Takeda, M., Muhammad Tunio, G., Matsuzawa, Y., Kanakura, Y., Shinomura, Y., & Kitamura, Y. (1998). Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science, 279*(5350), 577–580.

Ichikawa, H., Yoshida, A., Kanda, K., Kosugi, S., Ichikawa, T., Hanyu, T., Taguchi, T., Sakumoto, M., Katai, H., Kawai, A., Wakai, T., & Kondo, T. (2014). Prognostic significance of PML expression in gastrointestinal stromal tumor; integrated proteomic and transcriptomic analysis. *Cancer Science, 106*(1), 115–124.

Jiang, J., Jin, M. S., Suo, J., Wang, Y. P., He, L., & Cao, X. Y. (2012). Evaluation of malignancy using Ki-67, p53, EGFR and COX-2 expressions in gastrointestinal stromal tumors. *World Journal of Gastroenterology, 18*(20), 2569–2575. doi:10.3748/wjg.v18.i20.2569.

Joensuu, H. (2006). Gastrointestinal stromal tumor (GIST). *Annals of Oncology, 17*(Suppl 10), x280–x286. doi:10.1093/annonc/mdl274.

Joensuu, H. (2008). Risk stratification of patients diagnosed with gastrointestinal stromal tumor. *Human Pathology, 39*(10), 1411–1419. doi:10.1016/j.humpath.2008.06.025.

Joensuu, H., Vehtari, A., Riihimaki, J., Nishida, T., Steigen, S. E., Brabec, P., Plank, L., Nilsson, B., Cirilli, C., Braconi, C., Bordoni, A., Magnusson, M. K., Linke, Z., Sufliarsky, J., Federico, M., Jonasson, J. G., Dei Tos, A. P., & Rutkowski, P. (2012). Risk of recurrence of gastrointestinal stromal tumour after surgery: An analysis of pooled population-based cohorts. *The Lancet Oncology, 13*(3), 265–274. doi:10.1016/S1470-2045(11)70299-6.

Kang, H. J., Koh, K. H., Yang, E., You, K. T., Kim, H. J., Paik, Y. K., & Kim, H. (2006). Differentially expressed proteins in gastrointestinal stromal tumors with KIT and PDGFRA mutations. *Proteomics, 6*(4), 1151–1157. doi:10.1002/pmic.200500372.

Kang, R., Zhang, Q., Zeh, H. J., 3rd, Lotze, M. T., & Tang, D. (2013). HMGB1 in cancer: Good, bad, or both? *Clinical Cancer Research, 19*(15), 4046–4057. doi:10.1158/1078-0432.CCR-13-0495.

Kikuta, K., Gotoh, M., Kanda, T., Tochigi, N., Shimoda, T., Hasegawa, T., Katai, H., Shimada, Y., Suehara, Y., Kawai, A., Hirohashi, S., & Kondo, T. (2010). Pfetin as a prognostic biomarker in gastrointestinal stromal tumor: Novel monoclonal antibody and external validation study in multiple clinical facilities. *Japanese Journal of Clinical Oncology, 40*(1), 60–72. doi:hyp125 [pii]10.1093/jjco/hyp125.

Kikuta, K., Kubota, D., Saito, T., Orita, H., Yoshida, A., Tsuda, H., Suehara, Y., Katai, H., Shimada, Y., Toyama, Y., Sato, K., Yao, T., Kaneko, K., Beppu, Y., Murakami, Y., Kawai, A., & Kondo, T. (2012). Clinical proteomics identified ATP-dependent RNA helicase DDX39 as a novel biomarker to predict poor prognosis of patients with gastrointestinal stromal tumor. *Journal of Proteomics, 75*(4), 1089–1098. doi:10.1016/j.jprot.2011.10.005. S1874-3919(11)00491-X [pii].

Kindblom, L. G., Remotti, H. E., Aldenborg, F., & Meis-Kindblom, J. M. (1998). Gastrointestinal pacemaker cell tumor (GIPACT): Gastrointestinal stromal tumors show phenotypic characteristics of the interstitial cells of Cajal. *The American Journal of Pathology, 152*(5), 1259–1269.

Kondo, T., & Hirohashi, S. (2007). Application of highly sensitive fluorescent dyes (CyDye DIGE Fluor saturation dyes) to laser microdissection and two-dimensional difference gel electrophoresis (2D-DIGE) for cancer proteomics. *Nature Protocols, 1*(6), 2940–2956. doi:10.1038/nprot.2006.421.

Kondo, T., Kubota, D., & Kawai, A. (2012). Application of Proteomics to Soft Tissue Sarcomas. *International Journal of Proteomics, 2012*, 876401. doi:10.1155/2012/876401.

Kondo, T., Suehara, Y., Kikuta, K., Kubota, D., Tajima, T., Mukaihara, K., Ichikawa, H., & Kawai, A. (2013). Proteomic approach toward personalized sarcoma treatment: Lessons from prognostic biomarker discovery in gastrointestinal stromal tumor. *Proteomics Clinical Applications, 7*(1-2), 70–78. doi:10.1002/prca.201200085.

Kubota, D., Orita, H., Yoshida, A., Gotoh, M., Kanda, T., Tsuda, H., Hasegawa, T., Katai, H., Shimada, Y., Kaneko, K., Kawai, A., & Kondo, T. (2011). Pfetin as a prognostic biomarker for gastrointestinal stromal tumor: Validation study in multiple clinical facilities. *Japanese Journal of Clinical Oncology, 41*(10), 1194–1202. doi:hyr121 [pii]10.1093/jjco/hyr121.

Kubota, D., Okubo, T., Saito, T., Suehara, Y., Yoshida, A., Kikuta, K., Tsuda, H., Katai, H., Shimada, Y., Kaneko, K., Kawai, A., & Kondo, T. (2012). Validation study on pfetin and ATP-dependent RNA helicase DDX39 as prognostic biomarkers in gastrointestinal stromal yumour. *Japanese Journal of Clinical Oncology, 42*(8), 730–741. doi:hys092 [pii]10.1093/jjco/hys092.

Ladoire, S., Penault-Llorca, F., Senovilla, L., Dalban, C., Enot, D., Locher, C., Prada, N., Poirier-Colame, V., Chaba, K., Arnould, L., Ghiringhelli, F., Fumoleau, P., Spielmann, M., Delaloge, S., Poillot, M. L., Arveux, P., Goubar, A., Andre, F., Zitvogel, L., & Kroemer, G. (2015). Combined evaluation of LC3B puncta and HMGB1 expression predicts residual risk of relapse after adjuvant chemotherapy in breast cancer. *Autophagy, 11*(10), 1878–1890. doi:10.1080/15548627.2015.1082022.

Liegl, B., Kepten, I., Le, C., Zhu, M., Demetri, G. D., Heinrich, M. C., Fletcher, C. D., Corless, C. L., & Fletcher, J. A. (2008). Heterogeneity of kinase inhibitor resistance mechanisms in GIST. *The Journal of Pathology, 216*(1), 64–74. doi:10.1002/path.2382.

Linder, P., Lasko, P. F., Ashburner, M., Leroy, P., Nielsen, P. J., Nishi, K., Schnier, J., & Slonimski, P. P. (1989). Birth of the D-E-A-D box. *Nature, 337*(6203), 121–122. doi:10.1038/337121a0.

Linnekin, D., DeBerry, C. S., & Mou, S. (1997). Lyn associates with the juxtamembrane region of c-Kit and is activated by stem cell factor in hematopoietic cell lines and normal progenitor cells. *The Journal of Biological Chemistry, 272*(43), 27450–27455.

Lopes, L. F., & Bacchi, C. E. (2007). EGFR and gastrointestinal stromal tumor: An immunohistochemical and FISH study of 82 cases. *Modern Pathology, 20*(9), 990–994. doi:10.1038/modpathol.3800932.

Lotze, M. T., & DeMarco, R. A. (2003). Dealing with death: HMGB1 as a novel target for cancer therapy. *Current Opinion in Investigational Drugs, 4*(12), 1405–1409.

Melnick, A., & Licht, J. D. (1999). Deconstructing a disease: RARalpha, its fusion partners, and their roles in the pathogenesis of acute promyelocytic leukemia. *Blood, 93*(10), 3167–3215.

Miettinen, M., & Lasota, J. (2001). Gastrointestinal stromal tumors–definition, clinical, histological, immunohistochemical, and molecular genetic features and differential diagnosis. *Virchows Archiv, 438*(1), 1–12.

Miettinen, M., & Lasota, J. (2006). Gastrointestinal stromal tumors: Review on morphology, molecular pathology, prognosis, and differential diagnosis. *Archives of Pathology & Laboratory Medicine, 130*(10), 1466–1478. doi:10.1043/1543-2165(2006)130[1466:GSTROM]2.0.CO;2.

Miettinen, M., Lasota, J., & Sobin, L. H. (2005). Gastrointestinal stromal tumors of the stomach in children and young adults: A clinicopathologic, immunohistochemical, and molecular genetic study of 44 cases with long-term follow-up and review of the literature. *The American Journal of Surgical Pathology, 29*(10), 1373–1381.

Nagata, K., Kawakami, T., Kurata, Y., Kimura, Y., Suzuki, Y., Nagata, T., Sakuma, Y., Miyagi, Y., & Hirano, H. (2015). Augmentation of multiple protein kinase activities associated with secondary imatinib resistance in gastrointestinal stromal tumors as revealed by quantitative phosphoproteome analysis. *Journal of Proteomics, 115*, 132–142. doi:10.1016/j.jprot.2014.12.012.

Okamoto, Y., Sawaki, A., Ito, S., Nishida, T., Takahashi, T., Toyota, M., Suzuki, H., Shinomura, Y., Takeuchi, I., Shinjo, K., An, B., Ito, H., Yamao, K., Fujii, M., Murakami, H., Osada, H., Kataoka, H., Joh, T., Sekido, Y., & Kondo, Y. (2012). Aberrant DNA methylation associated with aggressiveness of gastrointestinal stromal tumour. *Gut, 61*(3), 392–401. doi:10.1136/gut.2011.241034.

Orita, H., Ito, T., Kushida, T., Sakurada, M., Maekawa, H., Wada, R., Suehara, Y., Kubota, D., & Sato, K. (2014). Pfetin as a risk factor of recurrence in gastrointestinal stromal tumors. *BioMedical Research International, 2014*, 651935. doi:10.1155/2014/651935.

Patel, S. R., Vadhan-Raj, S., Burgess, M. A., Plager, C., Papadopolous, N., Jenkins, J., & Benjamin, R. S. (1998). Results of two consecutive trials of dose-intensive chemotherapy with doxorubicin and ifosfamide in patients with sarcomas. *American Journal of Clinical Oncology, 21*(3), 317–321.

Patel, S. R., Gandhi, V., Jenkins, J., Papadopolous, N., Burgess, M. A., Plager, C., Plunkett, W., & Benjamin, R. S. (2001). Phase II clinical investigation of gemcitabine in advanced soft tissue sarcomas and window evaluation of dose rate on gemcitabine triphosphate accumulation. *Journal of Clinical Oncology, 19*(15), 3483–3489.

Prakash, S., Sarran, L., Socci, N., DeMatteo, R. P., Eisenstat, J., Greco, A. M., Maki, R. G., Wexler, L. H., LaQuaglia, M. P., Besmer, P., & Antonescu, C. R. (2005). Gastrointestinal stromal tumors in children and young adults: A clinicopathologic, molecular, and genomic study of 15 cases and review of the literature. *Journal of Pediatric Hematology/Oncology, 27*(4), 179–187.

Resendes, B. L., Kuo, S. F., Robertson, N. G., Giersch, A. B., Honrubia, D., Ohara, O., Adams, J. C., & Morton, C. C. (2004). Isolation from cochlea of a novel human intronless gene with predominant fetal expression. *Journal of the Association for Research in Otolaryngology, 5*(2), 185–202. doi:10.1007/s10162-003-4042-x.

Ridky, T. W., Chow, J. M., Wong, D. J., & Khavari, P. A. (2010). Invasive three-dimensional organotypic neoplasia from multiple normal human epithelia. *Nature Medicine, 16*(12), 1450–1455. doi:10.1038/nm.2265.

Rubin, B. P., Heinrich, M. C., & Corless, C. L. (2007). Gastrointestinal stromal tumour. *Lancet, 369*(9574), 1731–1741. doi:S0140-6736(07)60780-6 [pii]10.1016/S0140-6736(07)60780-6.

Salomoni, P., Ferguson, B. J., Wyllie, A. H., & Rich, T. (2008). New insights into the role of PML in tumour suppression. *Cell Research, 18*(6), 622–640. doi:10.1038/cr.2008.58.

Saponara, M., Urbini, M., Astolfi, A., Indio, V., Ercolani, G., Del Gaudio, M., Santini, D., Pirini, M. G., Fiorentino, M., Nannini, M., Lolli, C., Mandrioli, A., Gatto, L., Brandi, G., Biasco, G., Pinna, A. D., & Pantaleo, M. A. (2015). Molecular characterization of metastatic exon 11 mutant gastrointestinal stromal tumors (GIST) beyond KIT/PDGFRalpha genotype evaluated by next generation sequencing (NGS). *Oncotarget, 6*(39), 42243–42257. doi:10.18632/oncotarget.6278.

Sato, H., Hasegawa, T., Abe, Y., Sakai, H., & Hirohashi, S. (1999). Expression of E-cadherin in bone and soft tissue sarcomas: A possible role in epithelial differentiation. *Human Pathology, 30*(11), 1344–1349.

Scaife, C. L., Hunt, K. K., Patel, S. R., Benjamin, R. S., Burgess, M. A., Chen, L. L., Trent, J., Raymond, A. K., Cormier, J. N., Pisters, P. W., Pollock, R. E., & Feig, B. W. (2003). Is there a role for surgery in patients with "unresectable" cKIT+ gastrointestinal stromal tumors treated with imatinib mesylate? *The American Journal of Surgery, 186*(6), 665–669.

Serrels, A., McLeod, K., Canel, M., Kinnaird, A., Graham, K., Frame, M. C., & Brunton, V. G. (2012). The role of focal adhesion kinase catalytic activity on the proliferation and migration of squamous cell carcinoma cells. *International Journal of Cancer, 131*(2), 287–297. doi:10.1002/ijc.26351.

Shaw, J., Rowlinson, R., Nickson, J., Stone, T., Sweet, A., Williams, K., & Tonge, R. (2003). Evaluation of saturation labelling two-dimensional difference gel electrophoresis fluorescent dyes. *Proteomics, 3*(7), 1181–1195. doi:10.1002/pmic.200300439.

Shi, Z., Huang, Q., Chen, J., Yu, P., Wang, X., Qiu, H., Chen, Y., & Dong, Y. (2015). Correlation of HMGB1 expression to progression and poor prognosis of adenocarcinoma and squamous cell/adenosquamous carcinoma of gallbladder. *American Journal of Translational Research, 7*(10), 2015–2025.

Shrivastava, S., Mansure, J. J., Almajed, W., Cury, F., Ferbeyre, G., Popovic, M., Seuntjens, J., & Kassouf, W. (2015). The role of HMGB1 in radio-resistance of bladder cancer. *Molecular Cancer Therapeutics*. doi:10.1158/1535-7163.MCT-15-0581.

Silvestris, N., Parra, H. S., Angelini, F., Di Cosimo, S., D'Aprile, M., & Santoro, A. (2005). Lack of response to imatinib mesylate as second-line therapy in a patient with c-kit positive metastatic soft tissue leiomyosarcoma. *Tumori, 91*(1), 103.

Stiller, C. (2007). *Childhood cancer in Britian: Incidence, survival, mortality* (Vol. VII). New York: Oxford University Press.

Suehara, Y., Kondo, T., Fujii, K., Hasegawa, T., Kawai, A., Seki, K., Beppu, Y., Nishimura, T., Kurosawa, H., & Hirohashi, S. (2006). Proteomic signatures corresponding to histological classification and grading of soft-tissue sarcomas. *Proteomics, 6*(15), 4402–4409. doi:10.1002/pmic.200600196.

Suehara, Y., Kondo, T., Seki, K., Shibata, T., Fujii, K., Gotoh, M., Hasegawa, T., Shimada, Y., Sasako, M., Shimoda, T., Kurosawa, H., Beppu, Y., Kawai, A., & Hirohashi, S. (2008). Pfetin as a prognostic biomarker of gastrointestinal stromal tumors revealed by proteomics. *Clinical Cancer Research, 14*(6), 1707–1717. doi:14/6/1707 [pii]10.1158/1078-0432.CCR-07-1478.

Suehara, Y., Kikuta, K., Nakayama, R., Fujii, K., Ichikawa, H., Shibata, T., Seki, K., Hasegawa, T., Gotoh, M., Tochigi, N., Shimoda, T., Shimada, Y., Sano, T., Beppu, Y., Kurosawa, H., Hirohashi, S., Kawai, A., & Kondo, T. (2009). Anatomic site-specific proteomic signatures of gastrointestinal stromal tumors. *Proteomics – Clinical Applications, 3*(5), 584–596. doi:10.1002/prca.200800168.

Sugiura, T., Sakurai, K., & Nagano, Y. (2007). Intracellular characterization of DDX39, a novel growth-associated RNA helicase. *Experimental Cell Research, 313*(4), 782–790. doi:S0014-4827(06)00486-1 [pii]10.1016/j.yexcr.2006.11.014.

Taguchi, T., Sonobe, H., Toyonaga, S., Yamasaki, I., Shuin, T., Takano, A., Araki, K., Akimaru, K., & Yuri, K. (2002). Conventional and molecular cytogenetic characterization of a new human cell line, GIST-T1, established from gastrointestinal stromal tumor. *Laboratory Investigation, 82*(5), 663–665.

Takahashi, T., Serada, S., Ako, M., Fujimoto, M., Miyazaki, Y., Nakatsuka, R., Ikezoe, T., Yokoyama, A., Taguchi, T., Shimada, K., Kurokawa, Y., Yamasaki, M., Miyata, H., Nakajima, K., Takiguchi, S., Mori, M., Doki, Y., Naka, T., & Nishida, T. (2013). New findings of kinase switching in gastrointestinal stromal tumor under imatinib using phosphoproteomic analysis. *International Journal of Cancer, 133*(11), 2737–2743. doi:10.1002/ijc.28282.

Tang, D., Kang, R., Zeh, H. J., 3rd, & Lotze, M. T. (2010). High-mobility group box 1 and cancer. *Biochimica et Biophysica Acta, 1799*(1–2), 131–140. doi:10.1016/j.bbagrm.2009.11.014.

Timokhina, I., Kissel, H., Stella, G., & Besmer, P. (1998). Kit signaling through PI 3-kinase and Src kinase pathways: An essential role for Rac1 and JNK activation in mast cell proliferation. *The EMBO Journal, 17*(21), 6250–6262. doi:10.1093/emboj/17.21.6250.

Tran, T., Davila, J. A., & El-Serag, H. B. (2005). The epidemiology of malignant gastrointestinal stromal tumors: An analysis of 1,458 cases from 1992 to 2000. *The American Journal of Gastroenterology, 100*(1), 162–168. doi:10.1111/j.1572-0241.2005.40709.x.

Trent, J. C., Beach, J., Burgess, M. A., Papadopolous, N., Chen, L. L., Benjamin, R. S., & Patel, S. R. (2003). A two-arm phase II study of temozolomide in patients with advanced gastrointestinal stromal tumors and other soft tissue sarcomas. *Cancer, 98*(12), 2693–2699. doi:10.1002/cncr.11875.

Tuveson, D. A., Willis, N. A., Jacks, T., Griffin, J. D., Singer, S., Fletcher, C. D., Fletcher, J. A., & Demetri, G. D. (2001). STI571 inactivation of the gastrointestinal stromal tumor c-KIT oncoprotein: Biological and clinical implications. *Oncogene, 20*(36), 5054–5058. doi:10.1038/sj.onc.1204704.

Unlu, M., Morgan, M. E., & Minden, J. S. (1997). Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis, 18*(11), 2071–2077. doi:10.1002/elps.1150181133.

Vaira, V., Fedele, G., Pyne, S., Fasoli, E., Zadra, G., Bailey, D., Snyder, E., Faversani, A., Coggi, G., Flavin, R., Bosari, S., & Loda, M. (2010). Preclinical model of organotypic culture for pharmacodynamic profiling of human tumors. *Proceedings of the National Academy of Sciences of the United States of America, 107*(18), 8352–8356. doi:10.1073/pnas.0907676107.

Verweij, J., Casali, P. G., Zalcberg, J., LeCesne, A., Reichardt, P., Blay, J. Y., Issels, R., van Oosterom, A., Hogendoorn, P. C., Van Glabbeke, M., Bertulli, R., & Judson, I. (2004). Progression-free survival in gastrointestinal stromal tumours with high-dose imatinib: Randomised trial. *Lancet, 364*(9440), 1127–1134. doi:10.1016/S0140-6736(04)17098-0S0140673604170980 [pii].

Wang, F., Hansen, R. K., Radisky, D., Yoneda, T., Barcellos-Hoff, M. H., Petersen, O. W., Turley, E. A., & Bissell, M. J. (2002). Phenotypic reversion or death of cancer cells by altering signaling pathways in three-dimensional contexts. *Journal of the National Cancer Institute, 94*(19), 1494–1503.

Wardelmann, E., Thomas, N., Merkelbach-Bruse, S., Pauls, K., Speidel, N., Buttner, R., Bihl, H., Leutner, C. C., Heinicke, T., & Hohenberger, P. (2005). Acquired resistance to imatinib in gastrointestinal stromal tumours caused by multiple KIT mutations. *The Lancet Oncology, 6*(4), 249–251. doi:10.1016/S1470-2045(05)70097-8.

Yamada, K. M., & Cukierman, E. (2007). Modeling tissue morphogenesis and cancer in 3D. *Cell, 130*(4), 601–610. doi:10.1016/j.cell.2007.08.006.

Yamaguchi, U., Nakayama, R., Honda, K., Ichikawa, H., Hasegawa, T., Shitashige, M., Ono, M., Shoji, A., Sakuma, T., Kuwabara, H., Shimada, Y., Sasako, M., Shimoda, T., Kawai, A., Hirohashi, S., & Yamada, T. (2008). Distinct gene expression-defined classes of gastrointestinal stromal tumor. *Journal of Clinical Oncology, 26*(25), 4100–4108. doi:10.1200/JCO.2007.14.2331.

Yang, J., Eddy, J. A., Pan, Y., Hategan, A., Tabus, I., Wang, Y., Cogdell, D., Price, N. D., Pollock, R. E., Lazar, A. J., Hunt, K. K., Trent, J. C., & Zhang, W. (2010). Integrated proteomics and genomics analysis reveals a novel mesenchymal to epithelial reverting transition in leiomyosarcoma through regulation of slug. *Molecular & Cellular Proteomics, 9*(11), 2405–2413. doi:10.1074/mcp.M110.000240.

10 Proteogenomics for the Comprehensive Analysis of Human Cellular and Serum Antibody Repertoires

Paula Díez and Manuel Fuentes

Abstract

The vast repertoire of immunoglobulins produced by the immune system is a consequence of the huge amount of antigens to which we are exposed every day. The diversity of these immunoglobulins is due to different mechanisms (including VDJ recombination, somatic hypermutation, and antigen selection). Understanding how the immune system is capable of generating this diversity and which are the molecular bases of the composition of immunoglobulins are key challenges in the immunological field. During the last decades, several techniques have emerged as promising strategies to achieve these goals, but it is their combination which appears to be the fruitful solution for increasing the knowledge about human cellular and serum antibody repertoires.

In this chapter, we address the diverse strategies focused on the analysis of immunoglobulin repertoires as well as the characterization of the genomic and peptide sequences. Moreover, the advantages of combining various –omics approaches are discussed through review different published studies, showing the benefits in clinical areas.

Keywords

Antibody repertory • Immunoglobulin sequencing • Proteogenomics • Omics integration

P. Díez • M. Fuentes (✉)
Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), Avda. Universidad de Coimbra, S/N 37007 Salamanca, Spain

Proteomics Unit, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), Avda. Universidad de Coimbra, S/N 37007 Salamanca, Spain
e-mail: mfuentes@usal.es

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_10

10.1 Introduction

Since the landmark discovery of the antibody (also referred to as immunoglobulins, Ig) in blood serum – more than 100 years ago -, it has been well-characterized that a complex spectrum of distinct antibodies is contained in the blood. This wide spectrum is generated by B-cell clones through a somatic recombination process (Christiansen et al. 2007).

Recently, several studies have identified and determined the relative abundance of the monoclonal antibodies (mAbs) included in the serum pool that is elicited in response to vaccination or natural infection. To understand the humoral responses, it is important to obtain knowledge about the composition of the antigen-specific serum antibody repertoires, the properties (*e.g.*, affinity, recognized epitopes) of the respective Ig, the relationship between circulating Ig, and the presence of clonally expanded peripheral B-cells.

In the last years, the era of modern genomics and proteomics is providing extraordinary new tools for examining antibody repertoires. Next Generation Sequencing (NGS) allows the characterization of millions of B-cell receptor (BCR) sequences in a single experiment. Additionally, NGS approaches permit the study of the human antibody repertoire, not only to aid in the discovery of elite antibodies potentially useful as therapeutics, but also to comprehensively catalog the antibody sequences that are elicited during the adaptive immune response. Moreover, improvements made on NGS strategy have allowed the obtaining of the endogenous variable heavy and light (VH and VL) pairs within NGS datasets. Thus, thanks to the sequencing of this paired VH:VL, the BCR repertoire analysis has been successfully improved (Fig. 10.1). Finally, the combination of NGS approaches and high-resolution protein mass spectrometry (MS) has increased the characterization of serum antibodies (Lavinder et al. 2015).

Although deep sequencing technologies for DNA and RNA have been developed in the last years, this genomic and transcriptomic information does not exactly reflect the actual cellular state, as everything that is transcribed may not be translated (the translated portion of the genome is clearly reduced compared to the transcribed portion) (Haider and Pal 2013). Thus, proteogenomics appears as the combinatorial approach that employs proteomic information to supplement and increase the transcriptomic meaning (Woo et al. 2014).

10.2 Proteogenomics

Proteogenomics is an emerging field in which proteomics and genomics are integrated. The major goal of this recent discipline is to identify novel peptides, combining both strategies mentioned above. Thus, genomic and transcriptomic information is employed to generate customized protein sequence databases that are used as reference databases for MS data. In return, proteomics can offer protein-level evidence of gene expression (Castellana and Bafna 2010).

Thanks to the development of new sequencing approaches (*e.g.*, RNA-Seq, NGS) and the improvements made in the proteomics field, proteogenomics has experienced a significant increase in attention.

To understand the basis of proteogenomics, it is necessary to firstly describe the involved disciplines in the field (*i.e.*, genomics/transcriptomics and proteomics). Proteomics is the large-scale comprehensive study of proteins and looks for the characterization of their structures and functions, among other features. The termed was coined in 1994 by Marc Wilkins as a linguistic equivalent to the concept of genomics. There are different proteomic strategies to obtain proteomic data but generally 'shotgun proteomics' is the selected approach in which liquid chromatography (LC) and tandem mass spectrometry (MS/MS) are combined. To identify peptides with these approaches, it is necessary to use a reference database of theoretical protein sequences and it is in this point where a new strategy is required as many peptides are not present in any of the reference databases. Several alternatives have recently emerged, including sequence tag-based database searching or *de novo* sequencing (Nesvizhskii 2014). However, these approaches

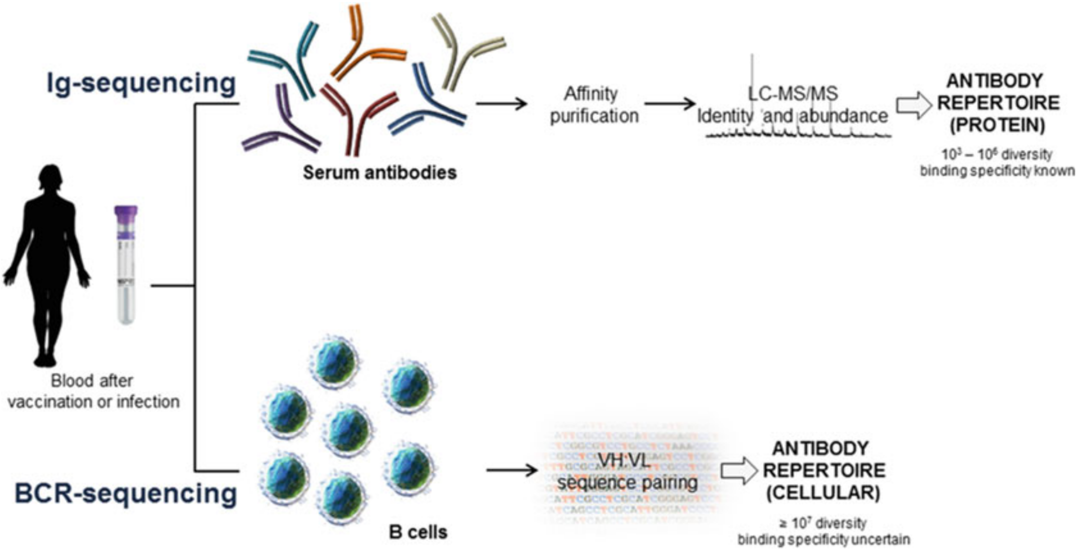


Fig. 10.1 Sequencing of immunoglobulins (Ig) and B-cell receptor (BCR). Schematic representation of the main approaches used to obtain the antibody repertoires (both protein and cellular) from serum antibodies and B-cells

are inefficient for large-scale studies since their running times are long and some reagents are too expensive (Fullwood et al. 2009; Hert et al. 2008). By these means, proteogenomics seems to be a good alternative to identify novel peptides. Nesvizhskii has suggested some guidelines for proteogenomic studies (Nesvizhskii 2014):

(i) Make available the customized protein sequence databases used to identify novel peptides
(ii) Query the peptides against all major reference databases and map to common sample contaminants
(iii) Describe the FDR estimation procedure
(iv) Mark peptides mapping to multiple genome locations

For peptide identification by proteogenomics approaches, it is crucial that the acquired MS/MS spectra are matched against a customized protein sequence database. There are diverse manners to develop such databases that we detail below.

- **Six-frame translation.** The sequence of nucleotides (DNA or RNA) can be split into consecutive and non-overlapping triplets and read in six different reading frames depending on the reading direction (5′ → 3′ or 3′ → 5′) and the starting reading point in the triplet (1st, 2nd, or 3rd nucleotide). When the genomic sequences of interest are translated in all 6 frames, the corresponding peptide / protein sequences for each frame are generated (Winnenburg et al. 2008). There are several computational strategies available to automatically generate these translations (*e.g.*, *getorf* from the EMBOSS (European Molecular Biology Open Software Suite), Bioline, ExPASy). This approach offers a huge diversity of possible peptides and the extremely large size of the resulting database constitutes a limitation. Moreover, most of the generated sequences do not really exist, requiring the establishment of rules (*e.g.*, homology to known coding sequences, minimum length) for the selection of the most likely frames (Nesvizhskii 2014).
- ***Ab initio* methods.** These approaches allow the prediction of genes – having no similarity to those previously described – with a high sensitivity and specificity. There is a wide range of informatics tools for *ab initio* gene prediction, such as EasyGene, GeneMark, MetaGene, Glimmer, Augustus, or GeneID

(Mathé et al. 2002). All these tools systematically look for *signals* and / or *content* revealing the presence of protein-coding sequences in the DNA. *Signals* are specific sequences that indicate the close presence of a gene (*e.g.*, Pribnow box), whereas *content* is referred to the statistics properties belonging to the coding sequence (*e.g.*, statistics for stop codons) (Zhu et al. 2010).

- **Expressed sequence tag (EST) data.** Short sequences (~300 nucleotides) resulting from a cDNA sequence are termed as expressed sequence tag (EST) and they represent portions of expressed genes. In public repositories, there is available a large number of ESTs generated by sequencing the 5′- or 3′- end of randomly isolated cDNA clones. There were 74,186,692 ESTs public entries registered in the EST database in the US National Center for Biotechnology Information (NCBI, January 11, 2016) (http://www.ncbi.nlm.nih.gov/genbank/dbest/dbest_summary/) (Teh et al. 2011). ESTs generate peptide sequence candidates in a more direct manner when compared to *ab initio* gene prediction approach, but a problem arises when using the six-frame translation of this data due to the resulting increase in the size of the generated database (Nesvizhskii 2014).
- **Annotated RNA transcripts.** Instead of generating protein sequences using the six-frame translation from DNA, there is an alternative using the three-frame translation of annotated RNA transcripts.
- **RNA-Seq data.** RNA-Seq can be considered as an improved version of ESTs in which the time and cost are significantly reduced when compared to Sanger sequencing. Moreover, its dynamic range is large thanks to the fact that sequencing reads mapping to unique regions are unlimited. However, the main inconvenience in this technique is the annotation of low abundance genes as they are represented by few reads (Roberts et al. 2011). This gene annotation approach has allowed the characterization of many transcriptomes, from specific organisms, such as *Deinococcus deserti* (de Groot et al. 2014), to organs involved in diseases (*e.g.*, the transcriptome of the brain endothelium from individuals with cerebrovascular dysfunction in ischemic stroke (Zhang et al. 2015)).

Once the protein sequence database has been customized, it is necessary to generate the proteomics data. The user can then identify the peptides by combining the two sources of information. One known strategy to increase the number of identified peptides is based on the application of multiple searching tools with the same dataset, together with rescoring of peptide identifications. Moreover, it is essential to estimate the identification confidence to avoid wrong identifications at the spectrum, peptide, and protein levels. With this aim, decoy sequences are included in the search database at each analysis level. With this information, a false discovery rate (FDR) can be calculated for the estimation of global error rates. Typically, it is assumed that FDR at peptide-spectrum matches (PSM), peptide, and protein levels should be ~1 % meaning that 99 % of the identifications do not match with decoy sequences from the search database and are considered as correct identifications. When dealing with novel peptides, it is necessary to take into consideration than they may present stronger evidence than known peptides to be assumed as novel peptides. Additionally, FDR estimation should be done separately for novel and known peptides (Reiter et al. 2009; Elias and Gygi 2010; Nesvizhskii 2014).

The integration of –omics strategies should provide new information about the behavior of the cell and the relation between the transcribed genes and their translated counterparts. Despite the promising results, there are still numerous challenges that should be addressed to increase the quality of the biological information. Several studies have recently dealt with this issue, showing encouraging outcomes as a result of the joining different ‘omics’ approaches. Below we discuss some of these studies.

The integration of proteomics and transcriptomics datasets was employed to analyze a lymphoma B-cell line (Ramos) by Díez and colleagues (Díez et al 2015). Proteomics data obtained using a nanoUPLC-LTQ-Orbitrap Velos was combined with the transcriptomic profiling of the Ramos cells revealing a 94 % overlap in the proteins identified by both 'omics' approaches. A further analysis showed 30 % coverage of all protein-coding genes present in the human genome. Original datasets were processed to classify the identified proteins in different groups depending on coverage and confidence levels. In this way, the generated datasets were as follows:

(i) *intersection*, proteins systematically identified in three replicated experiments with 2 or more proteotypic peptides at protein FDR < 0.01
(ii) *union*, proteins identified in any experiment with 2 or more proteotypic peptides at protein FDR < 0.01
(iii) *maximum*, proteins identified in any experiment with 1 or more proteotypic peptides at protein FDR < 0.01

This classification allows the selection of proteins in function of the stringency level and shows the complementarity between approaches to get a full map of the studied lymphoma B-cells.

McRedmond and colleagues (McRedmond et al. 2004) performed the integration of proteomics and genomics for the characterization of platelets. Specifically, they identified 82 secreted proteins and compared them to the transcriptome data obtaining 69 % correlation between both 'omics' data. Moreover, they predicted the presence of novel proteins in the platelets.

Another example is the study performed by Günther et al. in which they show that the combination of 'omics' strategies could be employed for dimension reduction and detection of candidate biomarkers in acute kidney transplant rejection (Günther et al. 2014).

10.3 Analysis Tools for Sequencing Immunoglobulin Genes

The immunoglobulins (Ig), also known as antibodies, are Y-shaped proteins, mainly produced by plasma cells (Fig. 10.2). Their function is to identify and neutralize antigens from pathogens. Igs are constituted by four chains: two identical light chains (L) and two identical heavy chains (H). In mammals, light chains can be classified into two types: lambda and kappa. These chains are connected by disulfide bonds at different points of the structure. Additionally, two parts can be distinguished in antibodies: the variable region (V) and the constant region (C). The variable region includes the epitope binding site where the antigen is recognized. Finally, the structure of the Ig can be divided into the Fab region, for the antigen specificity, and the Fc region, for determining the class effect of the antibody (Schroeder and Cavacini 2013).

The huge amount of different Ig proteins produced to counter the vast repertoire of antigens is generated thanks to a multi-layer mechanism developed by the immune system. The molecular sequence diversity of B and T cell receptors (BCR and TCR, respectively) determines their capacity to bind to a great diversity of antigens. This diversity is generated by three different but complementary mechanisms: VDJ recombination, somatic hypermutation, and antigen selection (Ralph and Matsen 2015).

The VDJ recombination process, discovered by Susumu Tonegawa (Hozumi and Tonegawa 1976), comprises the random selection of V, D, and J genes which will be further joined in a process that deletes some nucleotides from their sequences. The final sequence generated defines the specificity against the antigen (these sequences are known as complementary determining regions, CDR) (Ralph and Matsen 2015).

On the other hand, somatic hypermutation is a process in which single base substitutions are generated, with occasional deletions and insertions, in the variable region of the Ig, generated by the B-cells. The rate of somatic hypermutation in humans is 10^{-5}–10^{-3} mutations per base pair per generation. This process generates higher specific antibodies (Li et al. 2004; Kostareli et al. 2012).

Additionally, the antigen selection theory – developed by Burnet in the early 1940s – states that those B lymphocytes generating a specific antibody which is able to block a specific antigen

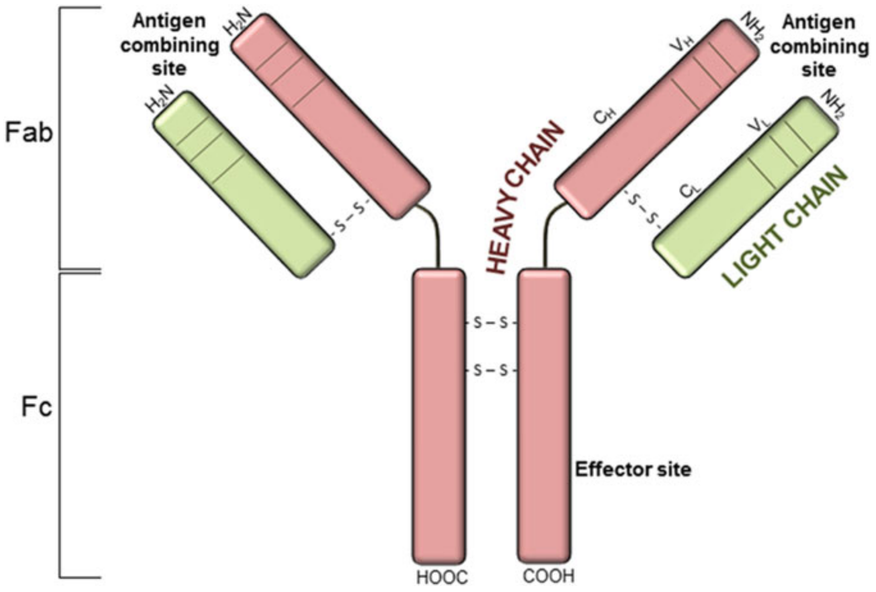


Fig. 10.2 A representative structure of an immunoglobulin. Four chains can be distinguished: two identical light chains and two identical heavy chains. Disulfide bonds connect both chains (–S–S–). C_H constant region of the heavy chain, V_H variable region of the heavy chain, C_L constant region of the light chain, V_L variable region of the light chain

are activated to generate clones for antibody production (Jordan and Baxter 2008).

With all these processes, the total diversity of generated Ig molecules is virtually unlimited. Then, their study should be focused on the detailed analysis of their sequences. With this purpose, the IgBLAST analysis tool was developed at NCBI. Its algorithm is based on the BLAST search algorithm including BCR-specific aspects:

(i) Reporting of the gene matches from the germline V, D, and J domains to the query sequence
(ii) Annotation of the Ig domain
(iii) Showing the V(D)J junction details
(iv) Information about the rearrangement (in-frame and out-of-frame)

Special requirements are needed for identifying Ig sequences, since they are quite long (290 bases for V genes). In this way, IgBLAST has the capacity to process multiple queries (<1000 sequences per batch) (Ye et al. 2009).

Additionally, there is an online annotation tool called IMGT, which is considered as the global reference in immunogenetics and immunoinformatics. Its content comprises information about Ig, TCR, and major histocompatibility complex (MHC), among others (Lefranc et al. 2015). Other examples of germline databases include VBASE2, IHMMune-align, and JoinSolver (Ye et al. 2009).

10.4 *De Novo* Protein Sequencing of Monoclonal Antibodies

Determining peptide amino acid sequences is possible through tandem MS. This method is known as "*de novo* peptide sequencing" and it has its beginnings in the Edman degradation procedure. Contrary to database searching, *de novo* sequencing allows the recognition of novel peptides as it assigns fragment ions from a mass spectrum (Hughes et al. 2010).

However, the application of *de novo* sequencing for identifying the sequences of antibodies is still a challenge. In these cases, MS/MS is not applicable and it makes it necessary to look for new approaches. Classical Edman degradation could be an option, but it is a low-throughput and time-consuming strategy. Nevertheless, Bandeira and colleagues have developed a new approach termed as Comparative Shotgun Protein Sequencing (CSPS) to identify unknown proteins using known proteins as templates in less than 72 h. The functioning of CSPS is based on a three-step sequence: alignment, assembly, and consensus approach. At the alignment step, spectral alignments from overlapping peptides are identified. The combination of spectral alignments into spectral contigs occurs at the assembly step, and these spectral contigs result in protein contigs at the consensus step (Bandeira et al. 2008).

As described in the Syd Labs web page (http://www.sydlabs.com/de-novo-antibody-sequencing-service-p58.htm), the general procedure for *de novo* antibody sequencing includes:

(i) Preliminary tryptic digestion followed by LC-MS/MS
(ii) Scale-up digestion with other enzymes (Asp-N, Glu-C) followed by fragmentation optimization
(iii) Separation of antibody heavy chain and light chain if required
(iv) Data analysis.

A large number of studies have applied *de novo* sequencing to determine antibody sequences. For instance, Pham and collaborators described the sequence of a full-length monoclonal antibody raised against OX40 ligand, including heavy and light chain sequences. In this study, they combined Edman degradation and mass spectrometric analysis (Pham et al. 2006).

In turn, Resemann and colleagues characterized the primary structure of a 13.6 kDa single heavy chain camelid antibody (V_HH) using top-down mass spectrometric analysis demonstrating that ~14 kDa proteins can be sequenced entirely by MS (Resemann et al. 2010).

Finally, the combination of high-resolution MS, *de novo* sequencing, and reverse engineering and chimerization approaches have allowed the purification and sequencing of antibodies derived for ascites, specifically, the heavy and light chain sequences of the LT-3 F12 antibody (Castellana et al. 2011).

10.5 Native Mass Spectrometry

Native mass spectrometry is an emerging technology for the investigation of the native-like quaternary structures. Although this approach does not offer molecular or atomic structure information, it is characterized by its numerous advantages, including high sensitivity, speed selectivity, unlimited and dynamic mass range, and accuracy. Moreover, it allows the isolation of a specific sample within a heterogeneous protein complex and the requirement of small amounts of sample (minimum of 10 picomoles) (van Duijn 2010). In summary, this tool can be considered as an intermediate platform between interactomics and structural biology, although other technologies are needed to cover the gaps (*e.g.*, nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, and electron microscopy). Native MS is also useful for refining structural models (Heck 2008; van Duijn 2010).

For native MS, electrospray ionization (ESI) is one of the most used options to prepare the sample for the further MS analysis. To preserve the quaternary protein structure, an aqueous ammonium acetate solution has been introduced, since it is compatible with the MS process. Regarding the mass analyzer, it is necessary that it has a high accuracy for ion identification (Heck 2008; van Duijn 2010).

Native MS has allowed the study of proteasomes, RNA polymerase II and III, and protein complexes, among others. Concerning proteasomes, several researchers have employed native MS to characterize them in different species (*e.g.*, *Methanosarcina thermophile*, *Thermoplasma acidophilum*, *Rhodococcus erythropolis*) providing information about their stoichiometries and masses (Heck 2008). Similar characterizations were done for RNA polymerases by Lorenzen and colleagues (Lorenzen et al. 2007) for polymerases from yeast.

Other studies have been focused on the analysis of monoclonal antibodies, including the evaluation of antibody-antigen binding, structural features, dynamics, and interaction strengths. Rosati and colleagues have characterized monoclonal antibodies and one of the first challenges that they found was related to the presence of N-linked glycosylation sites in each heavy chain as they are highly dependent of the type of cells. By applying native MS in orbitrap analyzers, they were able to perform the studies in a fast and sensitive manner (Rosati et al. 2012).

10.6 Clinical Applications of Proteogenomics Approaches

Bearing in mind all these advances in ProteoGenomics, the next step will focus on their applications in the clinical area, for both diagnostics and prognostics. For instance, improving the knowledge about the structural peculiarities of Igs may serve as a prediction factor of the possible development of pathologies, as was shown by Lomakin and colleagues (Lomakin et al. 2014) for multiple sclerosis disease. In this study, they looked for the "molecular signature" of viruses in the Ig repertoire, finding specific variable heavy and light chains from Ig against the myelin basic protein (MBP). Moreover, they revealed the cross-reactivity of this protein with the latent membrane protein 1 (LMP1) of Epstein-Barr virus which suggests that the responsible part of the antibody specificity is the light chain of the Ig.

Regarding cancer research, proteogenomics appears as a promising approach to better understand how cancer progression occurs. Thus far, the characterizations of molecular changes in cancer have been made by deep genome sequencing (led by the International Cancer Genome Consortium and The Cancer Genome Atlas, TCGA). However, the need of integrating genotypes and phenotypes, together with the development of proteomics, has accelerated the establishment of an international consortium (Clinical Proteomics Tumor Analysis Consortium, CPTAC) aimed at understanding the molecular basis of cancer (Faulkner et al. 2015). For instance, the proteogenomic characterization of human colon and rectal cancers has recently been described, which identified potential candidate biomarkers and therapeutic targets. These findings were made possible thanks to the association of global changes at mRNA and protein levels (Zhang et al. 2014). In turn, Fanayan et al. integrated proteomics (shotgun approach) and transcriptomics (RNA-Seq) approaches to evaluate human colon cancer cell lines (LIM1215, LIM1899, and LIM2405) within the chromosome-centric human proteome project (C-HPP). Their results included potential markers for colorectal cancer, including mortalin, nucleophosmin, ezrin, and exportin, among others (Fanayan et al. 2013).

10.7 Conclusions

In order to characterize the wide number of Igs produced by the immune system, it is required that new approaches are developed to achieve this goal. Traditionally, genome-scale techniques have been applied to increase the knowledge about Igs (*e.g.*, sequence of the antigen binding site, domains for epitope recognition). In this sense, numerous strategies have been successfully implemented in the genomics field. Special mention should be given to next generation sequencing (NGS) that allows characterizing millions of B-cell receptor sequences in a single experiment. Nevertheless, recent advances in the proteomics field have highlighted the need of new perspectives when analyzing antibody repertoires. Thus, combining strategies – not only classical genomics and transcriptomics techniques but also new 'omics' disciplines, such as proteomics and metabolomics – could actually signify an advance in the clinical area where immunoglobulins and their effects are crucial for the development of disease.

Acknowledgments We gratefully acknowledge financial support from the Carlos III Health Institute of Spain (ISCIII, FIS PI11/02114, FIS PI14/01538), Fondos

FEDER (EU) and Junta Castilla-León. BIO/SA07/15; Fundación Samuel Solórzano FS/23-2015. The Proteomics Unit belongs to ProteoRed, PRB2-ISCIII, FONDOS FEDER, supported by grant PT13/0001. P.D. is supported by a JCYL-EDU/346/2013 Ph.D. scholarship.

References

Bandeira, N., et al. (2008). Automated de novo protein sequencing of monoclonal antibodies. *Nature Biotechnology, 26*(12), 1336–1338.

Castellana, N., & Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of Proteomics, 73*(11), 2124–2135. Available at: http://dx.doi.org/10.1016/j.jprot.2010.06.007.

Castellana, N. E., et al. (2011). Resurrection of a clinical antibody: Template proteogenomic de novo proteomic sequencing and reverse engineering of an anti-lymphotoxin-alpha antibody. *Proteomics, 11*(3), 395–405.

Christiansen, M., et al. (2007). Chapter 4: Maritime transportation. In *Handbooks in operations research and management science*, 14(C) (pp. 189–284). Amsterdam: Elsevier.

Díez, P., et al. (2015). Integration of Proteomics and Transcriptomics data sets for the analysis of B-cell lymphoma cell line in the context of the Chromosome-Centric Human Proteome Project. *J. Proteome Res. 14*(9):3530–40.

de Groot, A., et al. (2014). RNA sequencing and proteogenomics reveal the importance of leaderless mRNAs in the radiation-tolerant bacterium Deinococcus deserti. *Genome Biology and Evolution, 6*(4), 932–948.

Elias, J. E., & Gygi, S. P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in Molecular Biology (Clifton, NJ), 604*(2), 55–71.

Fanayan, S., et al. (2013). Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. *Journal of Proteome Research, 12*, 1732–1742.

Faulkner, S., Dun, M. D., & Hondermarck, H. (2015). Proteogenomics: Emergence and promise. *Cellular and Molecular Life Sciences, 72*(5), 953–957. Available at: http://link.springer.com/10.1007/s00018-015-1837-y.

Fullwood, M. J., et al. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research, 19*(4), 521–532.

Günther, O. P., et al. (2014). Novel multivariate methods for integration of genomics and proteomics data: Applications in a kidney transplant rejection study. *Omics: A Journal of Integrative Biology, 18*(11), 682–695.

Haider, S., & Pal, R. (2013). Integrated analysis of transcriptomic and proteomic data. *Current Genomics, 14*(2), 91–110.

Heck, A. J. R. (2008). Native mass spectrometry: A bridge between interactomics and structural biology. *Nature Methods, 5*(11), 927–933.

Hert, D. G., Fredlake, C. P., & Barron, A. E. (2008). Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis, 29*, 4618–4626.

Hozumi, N., & Tonegawa, S. (1976). Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences of the United States of America, 73*(10), 3628–3632.

Hughes, C., Ma, B., & Lajoie, G. A. (2010). De novo sequencing methods in proteomics. *Methods in Molecular Biology (Clifton, NJ), 604*, 105–121.

Jordan, M. A., & Baxter, A. G. (2008). Quantitative and qualitative approaches to GOD: The first 10 years of the clonal selection theory. *Immunology and Cell Biology, 86*(1), 72–9.

Kostareli, E., et al. (2012). Immunoglobulin gene repertoire in chronic lymphocytic leukemia: Insight into antigen selection and microenvironmental interactions. *Mediterranean Journal of Hematology and Infectious Diseases, 4*(1), e2012052.

Lavinder, J. J., et al. (2015). Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Current Opinion in Chemical Biology, 24*, 112–120.

Lefranc, M.-P., et al. (2015). IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Research, 43*(Database issue), D413–D422.

Li, Z., et al. (2004). The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes and Development, 18*(1), 1–11.

Lomakin, Y. a., et al. (2014). Heavy-light chain interrelations of MS-associated immunoglobulins probed by deep sequencing and rational variation. *Molecular Immunology, 62*(2), 305–314.

Lorenzen, K., et al. (2007). Structural biology of RNA polymerase III: Mass spectrometry elucidates subcomplex architecture. *Structure (London, England: 1993), 15*(10), 1237–1245.

Mathé, C., et al. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research, 30*(19), 4103–4117.

McRedmond, J. P., et al. (2004). Integration of proteomics and genomics in platelets: A profile of platelet proteins and platelet-specific genes. *Molecular & Cellular Proteomics, 3*(2), 133–144.

Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods, 11*(11), 1114–1125. Available at: http://

www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4392723&tool=pmcentrez&rendertype=abstract.

Pham, V., et al. (2006). De novo proteomic sequencing of a monoclonal antibody raised against OX40 ligand. *Analytical Biochemistry, 352*(1), 77–86.

Ralph, D., & Matsen, F. A. (2015). Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation, *arXiv*(q-bio.PE),1–40.

Reiter, L., et al. (2009). Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & Cellular Proteomics, 8*(11), 2405–2417.

Resemann, A., et al. (2010). Top-down de novo protein sequencing of a 13.6 kDa camelid single heavy chain antibody by matrix-assisted laser desorption ionization-time-of-flight/time-of-flight mass spectrometry. *Analytical Chemistry, 82*(8), 3283–3292.

Roberts, A., et al. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics, 27*(17), 2325–2329.

Rosati, S., et al. (2012). Exploring an orbitrap analyzer for the characterization of intact antibodies by native mass spectrometry. *Angewandte Chemie – International Edition, 51*(52), 12992–12996.

Schroeder, H. W., & Cavacini, L. (2013). Structure and function of immunoglobulins. *Journal Allergy Clinical Immunology, 125*(125), S41–S52.

Teh, S.-L., et al. (2011). Development of expressed sequence tag resources for Vanda Mimi Palmer and data mining for EST-SSR. *Molecular Biology Reports, 38*(6), 3903–3909.

van Duijn, E. (2010). Current limitations in native mass spectrometry based structural biology. *Journal of the American Society for Mass Spectrometry, 21*(6), 971–978.

Winnenburg, R., et al. (2008). PHI-base update: Additions to the pathogen – Host interaction database. *Database, 36*(October 2007), 572–576.

Woo, S., et al. (2014). Proteogenomic database construction driven from large scale RNA-Seq data. *Journal of Proteome Research, 13*(1), 21–28.

Ye, X., Blonder, J., & Veenstra, T. D. (2009). Targeted proteomics for validation of biomarkers in clinical samples. *Briefings in Functional Genomics & Proteomics, 8*(2), 126–135.

Zhang, B., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature, 513*(7518), 382–387.

Zhang, J., et al. (2015). Altered long non-coding RNA transcriptomic profiles in brain microvascular endothelium after cerebral ischemia. *Experimental Neurology, 277*, 162–170.

Zhu, W., Lomsadze, A., & Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research, 38*(12), 1–15.

Antibody-Based Proteomics 11

Christer Wingren

Abstract

Antibody-based proteomic approaches play an important role in high-throughput, multiplexed protein expression profiling in health and disease. These antibody-based technologies will provide (miniaturized) set-ups capable of the simultaneously profiling of numerous proteins in a specific, sensitive, and rapid manner, targeting high- as well as low-abundant proteins, even in crude proteomes such as serum. The generated protein expression patterns, or proteomic snapshots, can then be transformed into proteomic maps, or detailed molecular fingerprints, revealing the composition of the target (sample) proteome at a molecular level. By using bioinformatics, candidate biomarker signatures can be deciphered and evaluated for clinical applicability. The approaches will provide unique opportunities for e.g. disease diagnostics, biomarker discovery, patient stratification, predicting disease recurrence, and evidence-based therapy selection. In this review, we describe the current status of the antibody-based proteomic approaches, focusing on antibody arrays. Furthermore, the current benefits and limitations of the approaches, as well as a set of selected key applications outlining the applicative potential will be discussed.

Keywords

Affinity proteomics • Antibody • Antibody arrays • Antibody-based proteomics • Biomarker • Disease proteomics • Protein expression profiling

11.1 Introduction

Mass spectrometry (MS) based approaches have so far constituted the main workhorse for protein expression profiling efforts (Ebhardt et al. 2015; Parker and Borchers 2014; Solier and Langen

C. Wingren (✉)
Department of Immunotechnology, Lund University,
Medicon Village, SE-223 81 Lund, Sweden
e-mail: christer.wingren@immun.lth.se

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6_11

2014). MS displays many advantages for this purpose, such as direct (absolute) identification, quantitative read-out possibilities, and suitability for hypothesis-free biomarker discovery. However, MS-based approaches are also associated with significant technical limitations, including sensitivity, resolution, accuracy, and reproducibility, especially when targeting complex samples, such as serum, where protein expression covers a huge dynamic range. The need for new proteomic technologies has been one of the main driving forces in the development of affinity proteomics, mainly represented by antibody-based approaches (Saerens et al. 2008; Uhlen and Ponten 2005; Voshol et al. 2009; Solier and Langen 2014; Borrebaeck and Wingren 2009a, 2014). Antibody-based proteomic approaches, such as antibody microarrays, have rapidly evolved from early proof-of-concept stages to high-performing proteome profiling assays, and today constitutes key established approaches within high-throughput (disease) proteomics (Borrebaeck and Wingren 2009a, 2014).

Antibody-based proteomics can thus be defined as the systematic generation and use of protein-specific antibodies to explore the proteome or parts thereof. The antibodies can be used for analysis of the specific protein targets in a wide range of assay platforms, as outlined in Table 11.1. Aiming for tissue protein profiling, candidate platforms could include immunohistochemistry (IHC), antibody-enriched selected reaction monitoring (SRM), global proteome survey (GPS), Triple-X, and reversed antibody microarrays, or reverse-phase protein microarrays (RPPA). When considering for biofluid protein expression profiling, candidate platforms could include ELISA, antibody-enriched-SRM, GPS, Triple-X, reverse antibody microarrays or RPPA, and antibody nano- and microarrays. The choice of platform will depend on the research question at hand (e.g. discovery study vs. validation study) and technical requirements (*e.g.*, sensitivity, throughput, and degree of multiplexity).

Table 11.1 Antibody-based proteomic approaches

Protein targets	Antibody-based proteomic approaches
Tissue protein profiling (e.g., tumor extracts)	Immunohistochemistry (IHC)
	Antibody-enriched selected reaction monitoring (SRM)
	Global proteome survey (GPS)
	Triple-X
	Reverse antibody microarrays, or reverse-phase protein microarrays (RPPA)
Biofluid protein profiling (e.g., serum)	ELISA
	Antibody-enriched selected reaction monitoring (SRM)
	Global proteome survey (GPS)
	Triple-X
	Reversed antibody microarrays, or RPPA
	Antibody nano- and microarray

11.2 Choice of Antibody

So far, antibodies are by far the most well-characterized and commonly used probe format within affinity proteomics, i.e. antibody-based proteomic approaches (Borrebaeck and Wingren 2011; Saerens et al. 2008; Solier and Langen 2014; Uhlen and Ponten 2005; Voshol et al. 2009). The antibodies will play a central role, acting as specific capture probes and the antibody format used will be essential, setting the stage for the technology (assay) platform. In more detail, the antibody format will directly or indirectly influence the:

(i) Performance of the probes in the selected technology platform
(ii) Range of specificities that can be generated and included
(iii) Supply/renewability of probes.

Hence, these three central aspects must be considered when selecting the antibody format/design. Here, we will briefly discuss the use of different antibody formats, including polyclonal antibodies (pAbs) vs. monoclonal antibodies (mAbs) vs. recombinant antibodies (recAbs). The use of antibodies vs. affinity reagents based on other scaffolds, such as affibodies (Renberg et al. 2005, 2007) and aptamers (Lao et al. 2009;

Walter et al. 2008; Cho et al. 2006; Collett et al. 2005), is outside the scope of this article, and has been reviewed elsewhere (Borrebaeck and Wingren 2007, 2009a; Wingren and Borrebaeck 2006; Wingren and Borrebaeck 2004).

pAbs display the advantage of multiple-epitope binding for the target protein, which makes them more suitable for cross-platform assays, potentially binding to both native and denatured forms of the antigen. However, the production of pAbs relies on immunization, and this probe format often shows a distinct lack of reproducibility upon re-immobilization with the same antigen, which makes this reagent less attractive as a renewable probe resource. While large-scale productions of pAbs have been successfully managed (Berglund et al. 2008; Uhlen and Hober 2009), this still poses a major logistical bottleneck. The pAB format has been successfully used in various antibody-based proteomic approaches, such as ELISA, IHC, Triple-X, and RPPA.

mAb preparations display a single-epitope specificity, making them highly attractive for specific applications. In fact, mAbs are currently the most commonly used immunoreagent in diagnostic applications (Borrebaeck 2000). However, the single-epitope specificity makes this reagent less useful across platforms, where the protein antigen might be partly denatured in different ways. The reagent is fully renewable, making it an attractive reagent, but the initial production of mAbs represents a key logistical bottleneck for large-scale efforts. mAbs have been successfully applied in, *e.g.*, IHC, antibody-enriched SRM, RPPA, and ELISA.

recAbs are often handled and selected using phage display technologies (Borrebaeck and Wingren 2011; Soderlind et al. 2000). Due to technical (size) limitations, the most commonly used antibody format is single-chain fragment variable (scFv) antibody, *i.e.*, the smallest fragment of an antibody still retaining its unique epitope-binding properties. These mono-specific reagents display many beneficial features, such as representing a renewable antibody source, the antibody library can be designed (engineered) on a molecular level to display desired features, such as on-chip stability in array-based applications (Borrebaeck and Wingren 2009a, 2011), they are produced without the use of animals, and they represents an attractive source towards generating antibodies against the entire proteome. Access to high-performing libraries and having the phage display technology established in the laboratory represents practical limitations. recAbs have been successfully used for, *e.g.*, antibody-enriched SRM, GPS, RPPA, ELISA, and in particular antibody nano- and microarrays (Borrebaeck and Wingren 2009a, 2011).

11.3 Antibody-Based Proteomics – Basic Technological Concepts and Considerations

Here, we describe the various antibody-based proteomic approaches used in brief, general terms, and we highlight their advantages and limitations (Table 11.2).

Immunohistochemistry (IHC) is a classical method to discover tissue biomarkers and translate them into routine clinical practice. This approach relies on antibodies to measure levels of the target proteins from formalin-fixed, paraffin embedded (FFPE) tissue slices. To increase the throughput, the set-up has been expanded from one tissue slice per slide to several tissue slices per slide, thus representing tissue microarrays (TMAs) (Table 11.2). For example, the TMA technology enabled up to 1000 FFPE tissue samples to be assembled in an array format (Braunschweig et al. 2004; Hewitt 2004). Hence, TMAs enables researchers to use a single slide to perform studies on large cohorts of tissues using only small amounts of reagents. IHC commonly relies on labelled antibodies for detection, often demanding visual inspection of each slice. Hence, standardization and automation have been central points for further technical developments in recent years. Key advantages are assay sensitivity and the fact that spatial resolution at cellular level can be accomplished, i.e. providing information about where the target protein is located. The latter can provide a deeper insight into normal

Table 11.2 Advantages and challenges of antibody-based technologies for tissue and/or biofluid protein expression profiling

Technology	Advantages	Challenges
IHC	Sensitivity	Specificity
	Spatial resolution at cellular level	Absolute quantification
	Works with FFPE tissue	
	Automated systems	
	Multiplexing	
Antibody-enriched SRM	Multiplexing	One antibody per target required
	Sensitivity	High instrument costs
	Specificity	Pre-defined targets (not designed for discovery)
	Quantitative	Complex sample preparation
		Throughput
GPS and Triple-X	Multiplexing	High instrument costs
	Sensitivity	Complex sample preparation
	Specificity	Throughput
	Quantitative	
	Discovery mode	
	One antibody per many targets	
ELISA	Sensitivity	Multiplexing
	Well-established in clinical laboratories	High sample consumption
	Specificity (sandwich approach)	
Reverse antibody arrays or RPPA	Multiplexing	Sensitivity
	Low reagent consumption	Specificity
	Broad sample compatibility	Semi-quantitative
	Low consumption of reagents	Few high-performing platforms at hand
	Sensitivity	
	Multiplexing	
	High throughput	

cellular functions and pathogenic mechanisms. The semi-denatured state of the sample proteins will place high demands on the antibody reagent in terms of specificity, to minimize both false-positive and false-negative results.

Combining the specific capture of the target by the antibody with the power of MS, *i.e.*, antibody-enriched SRM, paves the way for specific and sensitive detection and absolute quantification of proteins (Whiteaker et al. 2007, 2010) (Table 11.2). The antibodies are first used to capture and enrich the target proteins. The captured proteins are then eluted, digested and analyzed on tandem-MS. The MS set-up is pre-set to only look for selected target peptides. The sample could also be digested prior to the specific capture. Polyclonal as well as monoclonal antibodies have been used for capture. The set-up is limited by the fact that the targets are pre-defined and that one antibody per target is required. Hence, the platform is not designed for large-scale discovery efforts. But on the other hand, the set-up displays high specificity, adequate sensitivity, and can be multiplexed. The cost for the MS instrumentation is high. The set-up works for both tissue and biofluid protein expression profiling.

Recently, two similar novel concepts were presented, demonstrating one solution to how the combination of antibody capture and MS detection can be converted into a discovery set-up. The two concepts, were called Triple-X Proteomics (Poetz et al. 2009; Volk et al. 2012; Hoeppe et al. 2011) (TXP) and the Global Proteome Survey (Olsson et al. 2011, 2012a, b; Wingren et al. 2009) (GPS) and they are based on the same fundamental principle, and will provide unique opportunities to perform global proteomics in a species independent manner, using a very limited set of antibodies. Briefly, antibodies are generated against short peptide motifs, only four to six amino acid residues long, each motif being shared by 2–100 different proteins. These context independent motif specific antibodies could then be used to target motif containing peptides in a species independent manner. From a practical point of view, the proteome is digested, e.g. trypsinated, and the peptide-specific antibodies are then used to specifically capture and enrich motif-containing peptides. Next, the motif-containing peptides are detected and identified (sequenced) using tandem mass spectrometry, thereby enabling us to back-track the original proteins in a quantitative manner. By using only 200 motif-specific antibodies, each targeting a motif shared among 50 unique proteins, this would enable us to potentially target about half the non-redundant proteome. The GPS set-up is based on recAbs, while the Triple-X set-up relies on pAbs and/or Mabs. The platforms can be designed to provide absolute quantification, and are compatible with both tissue and biofluid protein expression profiling. The throughput, set by the MS step, represents a key limitation.

ELISA is currently the gold standards in clinical settings for measurements of proteins. The set-up is based on immobilizing the capture antibody, which specifically binds the target protein. A secondary antibody (sandwich set-up) is often used for detection of bound proteins. The set-up can deliver relative as well as absolute levels of the profiled proteins. pAbs and mAbs are the main antibody formats used. Highly specific and sensitive assays can be designed, and any sample format can be targeted as long as the protein (epitope) is accessible. The approach is limited by multiplexing and relatively high sample consumption.

The reverse antibody array, or RPPA, is a novel, miniaturized set-up providing several benefits (Nishizuka and Mills 2016; Voshol et al. 2009). In these set-ups, the sample is arrayed and the antibodies are added one by one to detect the target protein in each individual spot. Key advantages are multiplexing and low sample consumption. The platform enables large-scale screening of virtually any biological fluid, such as serum, urine, and saliva. In addition, tissue samples can also be profiled, provided that the proteins can be solubilized and arrayed. Dispensing low (pL range) volumes of complex samples will, however, limit the sensitivity of the assay. In more detail, the number of molecules of each individual protein adsorbed per spot will be a limiting factor in particular for low-abundant proteins. Hence, this assay set-up is more suitable for profiling medium- to high-abundant proteins.

The concept of antibody arrays is based on printing small volumes (pL scale) of numerous (a few to several hundreds) antibodies with the desired specificities on-by-one in an ordered pattern, an array (<1 cm^2), onto a solid support (Borrebaeck and Wingren 2009a, 2014). The arrayed antibodies will act as specific catcher molecules for the target proteins. These miniaturized arrays are incubated with μL-scale of crude, non-fractionated sample. Next, specifically bound analytes are detected and semi-quantified, mainly using fluorescence as a mode of detection (Wingren and Borrebaeck 2008). The complete assay is run within less than 4 h, where after the microarray images are transformed into protein expression profiles, or protein maps, revealing the detailed composition of the sample. Depending on the application at hand, different bioinformatic strategies can be applied (Borrebaeck and Wingren 2007, 2009b) to further explore the wealth of data generated, *e.g.*, pin-pointing differentially expressed protein analytes between, *e.g.*, disease patients and healthy controls (Bauer et al. 2006; Carlsson et al. 2011). The advantages of the technology are low consumption of reagents, multiplexing, sensitivity, and high throughput. The number of high-performing antibody array platforms is still low, most likely reflecting the complexity of developing such set-ups, which requires a truly multidisciplinary approach.

The antibody array is a relatively new proteomic technology that has been subject to intense development in recent years, going from proof-of-concept to established proteomic assays. The technology has been found to display a great potential for multiplexed protein expression profiling and biomarker discovery. The antibody array platforms are compatible with both tissue and biofluid protein expression profiling. Based on this, antibody arrays were selected as a showcase technology for antibody-based proteomic approaches and will be described in more detail below.

11.4 Antibody Nano- and Microarrays

The basic approach of generating miniaturized antibody arrays, ranging in size from mm^2 (nanoarrays, nm sized spot features) to cm^2 (microarrays, µm sized spot features) (Wingren and Borrebaeck 2007) is based on direct printing (Borrebaeck and Wingren 2007; Wingren and Borrebaeck 2007), self-addressing (Svedhem et al. 2003; Wacker and Niemeyer 2004; Wacker et al. 2004), or self-assembly (He et al. 2008a, 2008b; He and Taussig 2001; Ramachandran et al. 2004, 2006, 2008) of small amounts (femtomole range) of individual antibodies onto a solid support (Fig. 11.1). While planar arrays on solid microscope slides, such as plastic, glass, and silicon chips, constitute the dominating format, providing up to 16 sub-arrays per slide, multiplexed arrays have also been produced on the bottom of flat ELISA plate wells as well as on beads in solution, so called bead-arrays (Borrebaeck and Wingren 2009a; Schwenk et al. 2008; Wingren and Borrebaeck 2009; Wong et al. 2009). The array assay is run like a traditional ELISA, but consuming only µL scale volumes of the reagents and samples. It is noteworthy that complex, unfractionated proteomes, such as serum, plasma, urine, and tissue extracts, can, in contrast to many competing proteomic technologies, be directly used, meaning that the key issue of pre-fractionation of the sample is bypassed (Wingren and Borrebaeck 2009). Any sample format can be targeted, as long as the proteins are exposed/available (e.g. cell surface membrane proteins) and/or can be solubilized, including serum, plasma, urine, cerebrospinal fluid, intact cells, cell lysates, cell supernatants, and tissue extracts, etc. (Belov et al. 2001, 2003; Campbell et al. 2006; Dexlin et al. 2008; Dexlin-Mellby et al. 2010; Ingvarsson et al. 2007; Kristensson et al. 2012; Wingren et al. 2007; Alhamdani et al. 2010; Hoheisel et al. 2013). The samples are in most cases labeled with a fluorescent dye, either directly or indirectly, and interfaced with a fluorescent-based sensing (Kusnezow et al. 2007; Wingren and Borrebaeck 2008; Wingren et al. 2007). Label-free detection technologies have also been investigated, but additional technological developments will be required before they can be established and adapted, for review see (Borrebaeck and Wingren 2007, 2009a; Wingren and Borrebaeck 2006). These multiplexed assays display a dynamic four orders of magnitude or more, and assay sensitivities in the pM to fM range. This enables low-abundant (pg/ml)

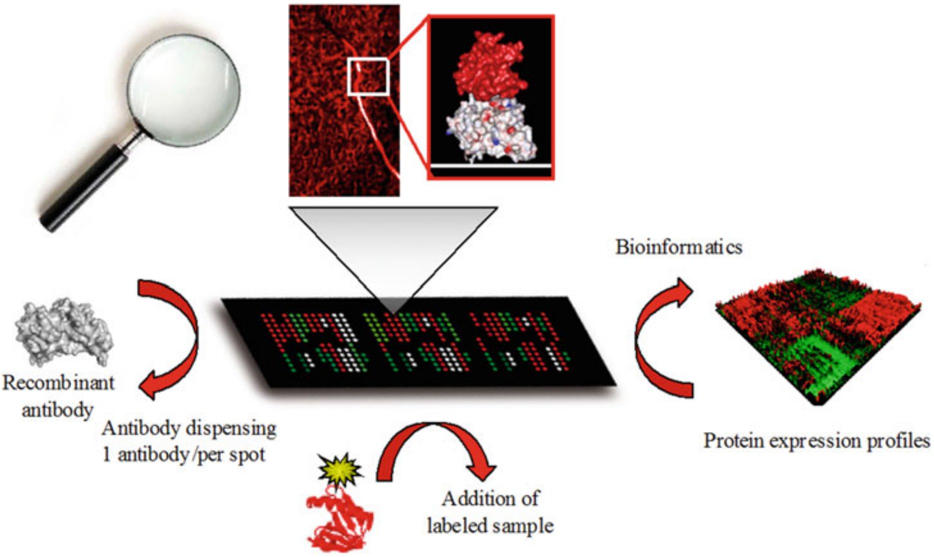


Fig. 11.1 Schematic illustration of the antibody microarray set-up

analytes to be directly profiled in crude proteomes. The assay time is similar to that of a conventional ELISA (about 4 h). By detecting and quantifying the signal intensity in each spot, the array images are transformed into protein expression profiles, deciphering the detailed composition of the sample. Finally, bioinformatics is applied to identify differences and similarities in protein expression profiles between the sample cohorts at hand, *e.g.*, cancer versus healthy controls, potentially generating candidate biomarker signatures. Typical applications of antibody-based microarrays include, but are not limited to, glycan profiling, delineation of signaling pathways, identification and detection of bacterial disease (proteins), cell surface membrane protein profiling of intact cells, as well as detection of disease associated biomarkers for diagnosis, prognosis, classification, evidence-based therapy selection, and predicting the risk for relapse (Alhamdani et al. 2010; Carlsson et al. 2010, 2011; Haab 2005; Sanchez-Carbayo et al. 2006; Wingren et al. 2012; Gao et al. 2005; Belov et al. 2001, 2003).

The process of designing, developing and applying antibody microarrays requires a cross-disciplinary approach to be adopted (Borrebaeck and Wingren 2009a). Consequently, five key basic principle areas needs to be addressed in a parallel manner, including:

(i) Antibody design
(ii) Array design
(iii) Sample handling
(iv) Assay design
(v) Data handling (bioinformatics)

Once these principles have been addressed and optimized, the technology is ready to be applied for the research problem at hand.

11.5 How Antibody Arrays Are Used Today in Research

Antibody microarrays are used to perform relative (or absolute) protein expression profiling of almost any kind of sample format, such as serum, often with the aim to decipher differentially expressed protein analytes and/or to delineate protein signatures for classification, for review see (Borrebaeck and Wingren 2007, 2009a, b; Haab 2005, 2006; Hartmann et al. 2009; Kingsmore 2006; Schwenk et al. 2008; Wingren and Borrebaeck 2009). The throughput per workstation per day varies, but can be in the range of hundred samples, each individual array assay in turn targeting anything from a few to several hundred protein analytes. However, the availability of high-performing antibody arrays, displaying the desired range of specificities, is in general a limiting factor. While a few groups have developed their own in-house antibody array set-ups (Haab and Zhou 2004; Hoheisel et al. 2013; Sanchez-Carbayo et al. 2006; Schroder et al. 2011; Schwenk et al. 2008; Wingren et al. 2007), other rely on commercially available alternatives, for review see (Borrebaeck and Wingren 2007, 2009a; Wingren and Borrebaeck 2009).

To date, a large number of antibody array-based applications have been presented, ranging from small proof-of-concept studies to large semi-global protein expression profiling studies (Table 11.3). As reviewing all antibody-array based applications to date is beyond the scope of this chapter, we have compiled a selected set of both early and more recent applications, giving a broad and representative view of what the technology can be used for. The compilation shows that the antibody array technology has been used in the following areas (Table 11.3)

1. Autoimmunity (Bauer et al. 2006, 2009; Carlsson et al. 2011; Szodoray et al. 2004; Lin et al. 2013; Kristensson et al. 2012)
2. Allergy (Lundberg et al. 2008)
3. Bladder proteomics (Fujita et al. 2006)
4. Cell proteomics (Campbell et al. 2006; De Ceuninck et al. 2004; Dexlin et al. 2008; Ko et al. 2005; Kopf et al. 2005; Tuomisto et al. 2005; Turtinen et al. 2004)
5. Drug abuse (Buechler et al. 1992)
6. Glycomics (Chen and Haab 2009; Chen et al. 2007; Yue et al. 2011)
7. Heart proteomics (Bereczki et al. 2007; Mitchell et al. 2005; Wu et al. 2004)

8. Hereditary disease (Srivastava et al. 2006; Jozwik et al. 2012)
9. Inflammatory conditions/infections (Madan et al. 2007; Kader et al. 2005; Cai et al. 2006; Sharma et al. 2006; Ingvarsson et al. 2007; Sandstrom et al. 2012)
10. Liver proteomics (Yee et al. 2007)
11. Lung proteomics (Izzotti et al. 2004)
12. Medical microbiology (Cai et al. 2005; Zhou et al. 2005, 2012; Gehring et al. 2008; Delehanty and Ligler 2002; Grow et al. 2003; Huang et al. 2003; Ligler et al. 2003; Rowe et al. 1999; Rowe-Taitt et al. 2000; Rubina et al. 2005; Taitt et al. 2002; Ellmark et al. 2006b; Anjum et al. 2006; Rucker et al. 2005)
13. Neurology/psychiatry (Kaukola et al. 2004; Sokolov and Cadet 2006; Krishnan et al. 2005)
14. Obstretics/gynaecology (Dexlin-Mellby et al. 2010; Wang et al. 2007; Centlow et al. 2011)
15. Oncoproteomics (Liu et al. 2011; Ahn et al. 2006; Sanchez-Carbayo et al. 2006; Carlsson et al. 2008, 2010, 2011; Celis et al. 2005; Hudelist et al. 2005; Lin et al. 2004; Orchekowski et al. 2005; Smith et al. 2006; Vazquez-Martin et al. 2007; Sreekumar et al. 2001; Ellmark et al. 2006a, b; Huang et al. 2001; Tannapfel et al. 2003; Belov et al. 2005, 2006; Zhou et al. 2004; Gao et al. 2005; Bartling et al. 2005; Ghobrial et al. 2005; Duffy et al. 2007; Mor et al. 2005; Ingvarsson et al. 2008; Schroder et al. 2010; Wingren et al. 2012; Miller et al. 2003; Shafer et al. 2007; Knezevic et al. 2001; Box et al. 2013; Sukhdeo et al. 2013; Yue et al. 2011; Patel et al. 2011; Sun et al. 2008; Hodgkinson et al. 2012; Shi et al. 2011; Ramirez and Lampe 2010; Yue et al. 2009)
16. Periodontology (Bodet et al. 2007)
17. Phosphoproteomics (Gembitsky et al. 2004; Flores-Delgado et al. 2007)
18. Protein expression (Han et al. 2006; Ivanov et al. 2004)
19. Protein signaling (Gaudet et al. 2005)

A majority of the applications have been performed within disease proteomics, and in particular oncoproteomics, but this does not reflect any limitation per se. In fact, as long as the target proteins can be addressed and the range of specificities of the arrayed antibodies is adequate for the application at hand, antibody arrays could be used for more or less any protein expression profiling application.

Using disease proteomics as a representative example, the project teams are frequently organized in a translational manner, involving scientists and clinicians with orthogonal competences, such as array technology, nanotechnology, protein engineering, immunochemistry, surface chemistry, sensing technology, bioinformatics, as well as disease biology, pathogenesis, and therapy (Borrebaeck and Wingren 2009a, b; Wingren and Borrebaeck 2009). The work is organized around a well-defined clinical problem, or set of problems, representing an unmet clinical need, and the project is frequently planned in a cross-disciplinary manner, going from bed-to-bench and back again. As for any proteomic study, it is essential that sequential studies are planned, going from discovery, pre-validation to validation studies, each step involving a new, independent patient data set to be targeted. In addition, the findings reported in each step of the project should also, if possible, be cross-validated using orthogonal methods (*e.g.*, ELISA and mass spectrometry).

11.6 Antibody Arrays – Selected Applications

As discussed above, we have compiled a selected set of both early and more recent antibody-array based applications, giving a representative view of what the technology can be used for (Table 11.3). The applications range from deciphering biomarker signatures for improved (and early) disease diagnosis, prognosis, predicting the risk for relapse, and evidence-based therapy selection, to detection and serotyping of bacteria. As a review of all of antibody array applications in detail is beyond the scope of this chapter, we

Table 11.3 Overview of selected antibody array-based applications

Area of application	Disease or biological process
Autoimmunity	Primary Sjögren´s syndrome
	Systemic lupus erythematosus
	Systemic sclerosis
Allergy	Cytokine profiling
Bladder proteomics	Smooth muscle hypertrophy
Cell proteomics	Amphotericin B exposure
	Blood phenotyping
	Cell differentiation
	Chondrocytes
	Model systems
Drug abuse	Screening
Glycomics	Pancreatic cancer
Heart proteomics	Myocardial infarction
Hereditary disease	Cystic fibrosis
Inflammation/infection	Artherosclerosis
	Inflammatory bowel disease
	Obesity
	Rhinovirus infection
	Complement deficiency
	Pancreatitis
Liver proteomics	APAP-induced liver disease
Lung proteomics	Chromium(VI)-treatment
Medical microbiology	Bacterial infection
	Detection of bacteria and/or toxins
	Helicobacter pylori infection
	Serotyping of bacteria
Neurology/psychiatry	Cerebral palsy
	Drug abuse
	Transverse myelitis
Obstetrics/gynaecology	Pre-eclampsia
Oncoproteomics	Angiogenesis
	Bladder cancer
	Breast cancer
	Colon cancer
	Colorectal cancer
	Gastric adenoma carcinoma
	Glioblastoma
	Hepatocellular carcinoma
	Leukemia
	Liver cancer
	Lung cancer
	Mantle-cell lymphoma
	Model system
	Ovarian cancer
	Pancreatic cancer
	Prostate cancer
	Squamous cell carcinoma
Periodontology	Model system
Phosphoproteomics	Model system
	Lung cancer
Protein expression	Post-translational modifications
	Biosynthetic pathways
Protein signaling	Proapoptotic/-survival stimuli

have chosen to focus on selected applications within disease proteomics, more specifically within the field of autoimmunity and cancer. To this end, we will display a few examples only as show cases to highlight the workflow and potential of the array methodology.

In the case of systemic lupus erythematosus (SLE), a chronic autoimmune connective tissue disease (Rovin and Zhang 2009; D'Cruz et al. 2007; Rahman and Isenberg 2008), the clinical need for serological/urinary biomarker signatures for improved diagnosis, prognosis, and classification is significant. In a discovery study by Carlsson et al., the authors showed that the first candidate serum biomarker signatures for diagnosis, prognosis, as well as sub-group phenotyping were successfully deciphered using recombinant antibody microarrays (Carlsson et al. 2011). Major efforts are currently under way to pre-validate and validate these promising findings, both enhancing our fundamental understanding of SLE and potentially paving the way for novel and improved clinical management of SLE patients (Wingren et al, unpublished observations).

In order to delineate a biomarker signature for bladder cancer, Sanchez-Carbayo et al. adopted a dual approach, combining the extraordinary power of both DNA microarrays and antibody microarrays (Sanchez-Carbayo et al. 2006). A set of candidate markers were first identified by gene profiling, after which an antibody microarray targeting a selected set of the candidate proteins was designed and applied. The data showed that the candidate biomarker signature discriminated between bladder cancer patients and healthy controls with a 94 % correct classification rate. The data also indicated a potential of stratifying the tumors (patients) into low versus high risk based on the overall survival of the bladder cancer patients.

Several array efforts have been devoted towards defining biomarkers for pancreatic cancer (Ingvarsson et al. 2008; Orchekowski et al. 2005; Schroder et al. 2010; Shi et al. 2011; Wingren et al. 2012; Yue et al. 2009, 2011; Gerdtsson et al. 2015). With an overall 5-year survival rate of less than 2–3 % pancreatic cancer is one of the most lethal types of malignancies (Chu et al. 2010; Jemal et al. 2009), which is why biomarkers for improved and early diagnosis would have a significant impact. Early work by Orchekowski et al. revealed a set of candidate serum biomarkers, but they proved to indicate on a general disease state rather than specifically pin-pointing pancreatic cancer. Interestingly, Yue and co-workers investigated the prevalence and nature of glycan alterations on specific proteins in pancreatic cancer patients using antibody-lectin sandwich arrays (Yue et al. 2009). Their work indicated a small set of significantly altered proteins that provided valuable insight into the prevalence and protein carriers of glycan alterations in pancreatic cancer. This outlines the potential of using glycan measurements on specific proteins for highly effective biomarkers. In three other studies, using recombinant antibody microarrays, candidate biomarkers for (early) diagnosis of pancreatic cancer have been deciphered (Ingvarsson et al. 2008; Wingren et al. 2012; Gerdtsson et al. 2015). Once validated, such biomarker signatures could pave the way for early and improved diagnosis based on a minimally invasive blood sample, which could result in a significantly improved outcome for pancreatic cancer patients. Shi and co-workers explored the possibility of defining potential markers for metastatic progression in pancreatic cancer using antibody microarrays, by comparing a metastatic pancreatic cancer line with its parental line (Shi et al. 2011). Interestingly, four dysregulated proteins were identified and validated, which might prove valuable for understanding pancreatic cancer metastasis and aid in the search for potential markers of metastatic progression.

11.7 Future Perspective

Antibody-based proteomic approaches will play a key role for high-throughput, multiplexed protein expression profiling in health and disease for years to come. This will enable simultaneous profiling of numerous high- and low-abundant proteins in crude sample formats in a highly selective, specific and sensitive manner, while consuming minimal amounts of reagents and sample. Generating high-resolution protein maps will be essential in the quest for deciphering biomarkers. In the end, this will pave the way for the next generation of disease diagnostics, patient stratification (*e.g.*, phenotyping, disease status, and sub-grouping), and predicting disease recurrence, as well as evidence-based therapy selection.

Acknowledgements This study was supported by grants from the Swedish National Science Council (VR-M and VR-NT) and VINNOVA.

References

Ahn, E. H., Kang, D. K., Chang, S. I., Kang, C. S., Han, M. H., & Kang, I. C. (2006). Profiling of differential protekin expression in angiogenin-induced HUVECs using antibody-arrayed ProteoChip. *Proteomics, 6*(4), 1104–1109.

Alhamdani, M. S., Schroder, C., & Hoheisel, J. D. (2010). Analysis conditions for proteomic profiling of mammalian tissue and cell extracts with antibody microarrays. *Proteomics, 10*(17), 3203–3207.

Anjum, M. F., Tucker, J. D., Sprigings, K. A., Woodward, M. J., & Ehricht, R. (2006). Use of miniaturized protein arrays for Escherichia coli O serotyping. *Clinical and Vaccine Immunology, 13*(5), 561–567.

Bartling, B., Hofmann, H. S., Boettger, T., Hansen, G., Burdach, S., Silber, R. E., & Simm, A. (2005). Comparative application of antibody and gene array for expression profiling in human squamous cell lung carcinoma. *Lung Cancer, 49*(2), 145–154.

Bauer, J. W., Baechler, E. C., Petri, M., Batliwalla, F. M., Crawford, D., Ortmann, W. A., Espe, K. J., Li, W., Patel, D. D., Gregersen, P. K., & Behrens, T. W. (2006). Elevated serum levels of interferon-regulated chemokines are biomarkers for active human systemic lupus erythematosus. *PLoS Medicine, 3*(12), e491.

Bauer, J. W., Petri, M., Batliwalla, F. M., Koeuth, T., Wilson, J., Slattery, C., Panoskaltsis-Mortari, A., Gregersen, P. K., Behrens, T. W., & Baechler, E. C. (2009). Interferon-regulated chemokines as biomarkers of systemic lupus erythematosus disease activity: A validation study. *Arthritis and Rheumatism, 60*(10), 3098–3107.

Belov, L., de la Vega, O., dos Remedios, C. G., Mulligan, S. P., & Christopherson, R. I. (2001). Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Research, 61*(11), 4483–4489.

Belov, L., Huang, P., Barber, N., Mulligan, S. P., & Christopherson, R. I. (2003). Identification of repertoires of surface antigens on leukemias using an antibody microarray. *Proteomics, 3*(11), 2147–2154.

Belov, L., Huang, P., Chrisp, J. S., Mulligan, S. P., & Christopherson, R. I. (2005). Screening microarrays of novel monoclonal antibodies for binding to T-, B- and myeloid leukaemia cells. *Journal Immunological Methods, 305*(1), 10–19.

Belov, L., Mulligan, S. P., Barber, N., Woolfson, A., Scott, M., Stoner, K., Chrisp, J. S., Sewell, W. A., Bradstock, K. F., Bendall, L., Pascovici, D. S., Thomas, M., Erber, W., Huang, P., Sartor, M., Young, G. A., Wiley, J. S., Juneja, S., Wierda, W. G., Green, A. R., Keating, M. J., & Christopherson, R. I. (2006). Analysis of human leukaemias and lymphomas using extensive immunophenotypes from an antibody microarray. *Brittish Journal of Haematology, 135*(2), 184–197.

Bereczki, E., Gonda, S., Csont, T., Korpos, E., Zvara, A., Ferdinandy, P., & Santha, M. (2007). Overexpression of biglycan in the heart of transgenic mice: An antibody microarray study. *Journal Proteome Research, 6*(2), 854–861.

Berglund, L., Bjorling, E., Oksvold, P., Fagerberg, L., Asplund, A., Szigyarto, C. A., Persson, A., Ottosson, J., Wernerus, H., Nilsson, P., Lundberg, E., Sivertsson, A., Navani, S., Wester, K., Kampf, C., Hober, S., Ponten, F., & Uhlen, M. (2008). A genecentric Human Protein Atlas for expression profiles based on antibodies. *Molecular & Cellular Proteomics, 7*(10), 2019–2027.

Bodet, C., Andrian, E., Tanabe, S. I., & Grenier, D. (2007). Actinobacillus actinomycetemcomitans lipopolysaccharide regulates matrix metalloproteinase, tissue inhibitors of matrix metalloproteinase, and plasminogen activator production by human gingival fibroblasts: A potential role in connective tissue destruction. *Journal of Cellular Physiology, 212*(1), 189–194.

Borrebaeck, C. A. K. (2000). Antibodies in diagnostics – From immunoassays to protein chips. *Immunology Today, 21*(8), 379–382.

Borrebaeck, C. A. K., & Wingren, C. (2007). High-throughput proteomics using antibody microarrays: An update. *Expert Review of Molecular Diagnostics, 7*(5), 673–686.

Borrebaeck, C. A., & Wingren, C. (2009a). Design of high-density antibody microarrays for disease proteomics: Key technological issues. *Journal of Proteomics, 72*(6), 928–935.

Borrebaeck, C. A. K., & Wingren, C. (2009b). Transferring proteomic discoveries into clinical practice. *Expert Review of Proteomics, 6*(1), 11–13.

Borrebaeck, C. A. K., & Wingren, C. (2011). Recombinant antibodies for the generation of antibody arrays. In U. Korf (Ed.), *Protein microarrays. methods in molecular biology. (Clifton, NJ), 785*, 247–262.

Borrebaeck, C. A. K., & Wingren, C. (2014). Antibody array generation and use. *Methods in Molecular Biology (Clifton, NJ), 1131*, 563–571.

Box, C., Zimmermann, M., & Eccles, S. (2013). Molecular markers of response and resistance to EGFR inhibitors in head and neck cancers. *Frontiers in Bioscience, 18*, 520–542.

Braunschweig, T., Chung, J. Y., & Hewitt, S. M. (2004). Perspectives in tissue microarrays. *Combinatorial Chemistry & High Throughput Screening, 7*(6), 575–585.

Buechler, K. F., Moi, S., Noar, B., McGrath, D., Villela, J., Clancy, M., Shenhav, A., Colleymore, A., Valkirs, G., Lee, T., et al. (1992). Simultaneous detection of seven drugs of abuse by the Triage panel for drugs of abuse. *Clinical Chemistry, 38*(9), 1678–1684.

Cai, H. Y., Lu, L., Muckle, C. A., Prescott, J. F., & Chen, S. (2005). Development of a novel protein microarray method for serotyping Salmonella enterica strains. *Journal Clinical Microbiology, 43*(7), 3427–3430.

Cai, M., Yin, W., Li, Q., Liao, D., Tsutsumi, K., Hou, H., Liu, Y., Zhang, C., Li, J., Wang, Z., & Xiao, J. (2006). Effects of NO-1886 on inflammation-associated cytokines in high-fat/high-sucrose/high-cholesterol diet-fed miniature pigs. *European Journal Pharmacology, 540*(1–3), 139–146.

Campbell, C. J., O'Looney, N., Chong Kwan, M., Robb, J. S., Ross, A. J., Beattie, J. S., Petrik, J., & Ghazal, P. (2006). Cell interaction microarray for blood phenotyping. *Analytical Chemistry, 78*(6), 1930–1938.

Carlsson, A., Wingren, C., Ingvarsson, J., Ellmark, P., Baldertorp, B., Ferno, M., Olsson, H., & Borrebaeck, C. A. K. (2008). Serum proteome profiling of metastatic breast cancer using recombinant antibody microarrays. *European Journal of Cancer, 44*(3), 472–480.

Carlsson, A., Persson, O., Ingvarsson, J., Widegren, B., Salford, L., Borrebaeck, C. A. K., & Wingren, C. (2010). Plasma proteome profiling reveals biomarker patterns associated with prognosis and therapy selection in glioblastoma multiforme patients. *Proteomics Clinical Applications, 4*(6–7), 591–602.

Carlsson, A., Wuttge, D. M., Ingvarsson, J., Bengtsson, A. A., Sturfelt, G., Borrebaeck, C. A. K., & Wingren, C. (2011). Serum protein profiling of systemic lupus erythematosus and systemic sclerosis using recombinant antibody microarrays. *Molecular & Cellular Proteomics 10*(5), M110 005033.

Celis, J. E., Moreira, J. M., Cabezon, T., Gromov, P., Friis, E., Rank, F., & Gromova, I. (2005). Identification of extracellular and intracellular signaling components of the mammary adipose tissue and its interstitial fluid in high risk breast cancer patients: Toward dissecting the molecular circuitry of epithelial-adipocyte stromal cell interactions. *Molecular & Cellular Proteomics, 4*(4), 492–522.

Centlow, M., Wingren, C., Borrebaeck, C. A. K., Brownstein, M. J., & Hansson, S. R. (2011). Differential gene expression analysis of placentas with increased vascular resistance and pre-eclampsia using whole-genome microarrays. *Journal of Pregnancy, 2011*, 472354.

Chen, S., & Haab, B. B. (2009). Analysis of glycans on serum proteins using antibody microarrays. *Methods of Molecular Biology, 520*, 39–58.

Chen, S., Laroche, T., Hamelinck, D., Bergsma, D., Brenner, D., Simeone, D., Brand, R. E., & Haab, B. B. (2007). Multiplexed analysis of glycan variation on native proteins captured by antibody microarrays. *Nature Methods, 4*(5), 437–444.

Cho, E. J., Collett, J. R., Szafranska, A. E., & Ellington, A. D. (2006). Optimization of aptamer microarray technology for multiple protein targets. *Analytica Chimica Acta, 564*(1), 82–90.

Chu, D., Kohlmann, W., & Adler, D. G. (2010). Identification and screening of individuals at increased risk for pancreatic cancer with emphasis on known environmental and genetic factors and hereditary syndromes. *Journal of the Pancreas, 11*(3), 203–212.

Collett, J. R., Cho, E. J., Lee, J. F., Levy, M., Hood, A. J., Wan, C., & Ellington, A. D. (2005). Functional RNA microarrays for high-throughput screening of antiprotein aptamers. *Analytical Biochemistry, 338*(1), 113–123.

D'Cruz, D. P., Khamashta, M. A., & Hughes, G. R. (2007). Systemic lupus erythematosus. *Lancet, 369*(9561), 587–596.

De Ceuninck, F., Dassencourt, L., & Anract, P. (2004). The inflammatory side of human chondrocytes unveiled by antibody microarrays. *Biochemical and Biophysical Research Communications, 323*(3), 960–969.

Delehanty, J. B., & Ligler, F. S. (2002). A microarray immunoassay for simultaneous detection of proteins and bacteria. *Analytical Chemistry, 74*(21), 5681–5687.

Dexlin, L., Ingvarsson, J., Frendeus, B., Borrebaeck, C. A. K., & Wingren, C. (2008). Design of recombiant antibody microarrays for cell surface membrane proteomics. *Journal of Proteome Research, 7*(1), 319–327.

Dexlin-Mellby, L., Sandström, A., Centlow, M., Nygren, S., Hansson, S. R., Borrebaeck, C. A. K., & Wingren, C. (2010). Tissue protome profiling of preeclamptic placenta tissue using recombinant antibody microarrays. *Proteomics – Clinical Applications, 4*(10–11), 794–807.

Duffy, H. S., Iacobas, I., Hotchkiss, K., Hirst-Jensen, B. J., Bosco, A., Dandachi, N., Dermietzel, R., Sorgen, P. L., & Spray, D. C. (2007). The gap junction protein connexin32 interacts with the Src homology 3/Hook domain of discs large homolog 1. *Journal Biological Chemistry, 282*(13), 9789–9796.

Ebhardt, H. A., Root, A., Sander, C., & Aebersold, R. (2015). Applications of targeted proteomics in systems biology and translational medicine. *Proteomics, 15*(18), 3193–3208.

Ellmark, P., Belov, L., Huang, P., Lee, C. S., Solomon, M. J., Morgan, D. K., & Christopherson, R. I. (2006a). Multiplex detection of surface molecules on colorectal cancers. *Proteomics, 6*(6), 1791–1802.

Ellmark, P., Ingvarsson, J., Carlsson, A., Lundin, B. S., Wingren, C., & Borrebaeck, C. A. K. (2006b). Identification of protein expression signatures associated with Helicobacter pylori infection and gastric adenocarcinoma using recombinant antibody microarrays. *Molecular & Cellular Proteomics, 5*(9), 1638–1646.

Flores-Delgado, G., Liu, C. W., Sposto, R., & Berndt, N. (2007). A limited screen for protein interactions reveals new roles for protein phosphatase 1 in cell cycle control and apoptosis. *Journal Proteome Research, 6*(3), 1165–1175.

Fujita, O., Asanuma, M., Yokoyama, T., Miyazaki, I., Ogawa, N., & Kumon, H. (2006). Involvement of STAT3 in bladder smooth muscle hypertrophy following bladder outlet obstruction. *Acta Medica Okayama, 60*(6), 299–309.

Gao, W. M., Kuick, R., Orchekowski, R. P., Misek, D. E., Qiu, J., Greenberg, A. K., Rom, W. N., Brenner, D. E., Omenn, G. S., Haab, B. B., & Hanash, S. M. (2005).

Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. *BMC Cancer, 5*, 110.

Gaudet, S., Janes, K. A., Albeck, J. G., Pace, E. A., Lauffenburger, D. A., & Sorger, P. K. (2005). A compendium of signals and responses triggered by prodeath and prosurvival cytokines. *Molecular & Cellular Proteomics, 4*(10), 1569–1590.

Gehring, A. G., Albin, D. M., Reed, S. A., Tu, S. I., & Brewster, J. D. (2008). An antibody microarray, in multiwell plate format, for multiplex screening of foodborne pathogenic bacteria and biomolecules. *Analytical Bioanalytical Chemistry, 39*(1), 497–506.

Gembitsky, D. S., Lawlor, K., Jacovina, A., Yaneva, M., & Tempst, P. (2004). A prototype antibody microarray platform to monitor changes in protein tyrosine phosphorylation. *Molecualr & Cellular Proteomics, 3*(11), 1102–1118.

Gerdtsson, A. S., Malats, N., Sall, A., Real, F. X., Porta, M., Skoog, P., Persson, H., Wingren, C., & Borrebaeck, C. A. K. (2015). A multicenter trial defining a serum protein signature associated with pancreatic ductal adenocarcinoma. *International Journal Proteomics, 2015*, 587250.

Ghobrial, I. M., McCormick, D. J., Kaufmann, S. H., Leontovich, A. A., Loegering, D. A., Dai, N. T., Krajnik, K. L., Stenson, M. J., Melhem, M. F., Novak, A. J., Ansell, S. M., & Witzig, T. E. (2005). Proteomic analysis of mantle-cell lymphoma by protein microarray. *Blood, 105*(9), 3722–3730.

Grow, A. E., Wood, L. L., Claycomb, J. L., & Thompson, P. A. (2003). New biochip technology for label-free detection of pathogens and their toxins. *Journal of Microbiology Methods, 53*(2), 221–233.

Haab, B. B. (2005). Antibody arrays in cancer research. *Molecular & Cellular Proteomics, 4*(4), 377–383.

Haab, B. B. (2006). Applications of antibody array platforms. *Current Opinion in Biotechnology, 17*(4), 415–421.

Haab, B. B., & Zhou, H. (2004). Multiplexed protein analysis using spotted antibody microarrays. *Methods in Molecular Biology (Clifton, NJ), 264*, 33–45.

Han, M. K., Hong, M. Y., Lee, D., Lee, D. E., Noh, G. Y., Lee, J. H., Kim, S. H., & Kim, H. S. (2006). Expression profiling of proteins in L-threonine biosynthetic pathway of Escherichia coli by using antibody microarray. *Proteomics, 6*(22), 5929–5940.

Hartmann, M., Roeraade, J., Stoll, D., Templin, M. F., & Joos, T. O. (2009). Protein microarrays for diagnostic assays. *Analytical and Bioanalytical Chemistry, 393*(5), 1407–1416.

He, M., & Taussig, M. J. (2001). Single step generation of protein arrays from DNA by cell-free expression and in situ immobilisation (PISA method). *Nucleic Acids Research, 29*(15), E73–73.

He, M., Stoevesandt, O., Palmer, E. A., Khan, F., Ericsson, O., & Taussig, M. J. (2008a). Printing protein arrays from DNA arrays. *Nature Methods, 5*(2), 175–177.

He, M., Stoevesandt, O., & Taussig, M. J. (2008b). In situ synthesis of protein arrays. *Current Opinion in Biotechnology, 19*(1), 4–9.

Hewitt, S. M. (2004). Design, construction, and use of tissue microarrays. *Methods in Molecular Biology (Clifton, NJ), 264*, 61–72.

Hodgkinson, V. C., EL, D., Agarwal, V., Garimella, V., Russell, C., Long, E. D., Fox, J. N., McManus, P. L., Mahapatra, T. K., Kneeshaw, P. J., Drew, P. J., Lind, M. J., & Cawkwell, L. (2012). Proteomic identification of predictive biomarkers of resistance to neoadjuvant chemotherapy in luminal breast cancer: a possible role for 14-3-3 theta/tau and tBID? *Journal of Proteomics, 75*(4), 1276–1283.

Hoeppe, S., Schreiber, T. D., Planatscher, H., Zell, A., Templin, M. F., Stoll, D., Joos, T. O., & Poetz, O. (2011). Targeting peptide termini, a novel immunoaffinity approach to reduce complexity in mass spectrometric protein identification. *Molecular & Cellular Proteomics, 10*(2), M110 002857.

Hoheisel, J. D., Alhamdani, M. S., & Schroder, C. (2013). Affinity-based microarrays for proteomic analysis of cancer tissues. *Proteomics Clinical Application, 7*(1–2), 8–15.

Huang, R. P., Huang, R., Fan, Y., & Lin, Y. (2001). Simultaneous detection of multiple cytokines from conditioned media and patient's sera by an antibody-based protein array system. *Analytical Biochemistry, 294*(1), 55–62.

Huang, T. T., Sturgis, J., Gomez, R., Geng, T., Bashir, R., Bhunia, A. K., Robinson, J. P., & Ladisch, M. R. (2003). Composite surface for blocking bacterial adsorption on protein biochips. *Biotechnology Bioenginering, 81*(5), 618–624.

Hudelist, G., Singer, C. F., Kubista, E., & Czerwenka, K. (2005). Use of high-throughput arrays for profiling differentially expressed proteins in normal and malignant tissues. *Anti-Cancer Drugs, 16*(7), 683–689.

Ingvarsson, J., Larsson, A., Sjoholm, A. G., Truedsson, L., Jansson, B., Borrebaeck, C. A. K., & Wingren, C. (2007). Design of recombinant antibody microarrays for serum protein profiling: Targeting of complement proteins. *Journal of Proteome Research, 6*(9), 3527–3536.

Ingvarsson, J., Wingren, C., Carlsson, A., Ellmark, P., Wahren, B., Engstrom, G., Harmenberg, U., Krogh, M., Peterson, C., & Borrebaeck, C. A. K. (2008). Detection of pancreatic cancer using antibody microarray-based serum protein profiling. *Proteomics, 8*(11), 2211–2219.

Ivanov, S. S., Chung, A. S., Yuan, Z. L., Guan, Y. J., Sachs, K. V., Reichner, J. S., & Chin, Y. E. (2004). Antibodies immobilized as arrays to profile protein post-translational modifications in mammalian cells. *Molecular & Cellular Proteomics, 3*(8), 788–795.

Izzotti, A., Bagnasco, M., Cartiglia, C., Longobardi, M., & De Flora, S. (2004). Proteomic analysis as related to transcriptome data in the lung of chromium(VI)-

treated rats. *International Journal of Oncology, 24*(6), 1513–1522.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., & Thun, M. J. (2009). Cancer statistics, 2009. *CA Cancer Journal for Clinicians, 59*(4), 225–249.

Jozwik, C. E., Pollard, H. B., Srivastava, M., Eidelman, O., Fan, Q., Darling, T. N., & Zeitlin, P. L. (2012). Antibody microarrays: analysis of cystic fibrosis. *Methods Molecular Biology, 823*, 179–200.

Kader, H. A., Tchernev, V. T., Satyaraj, E., Lejnine, S., Kotler, G., Kingsmore, S. F., & Patel, D. D. (2005). Protein microarray analysis of disease activity in pediatric inflammatory bowel disease demonstrates elevated serum PLGF, IL-7, TGF-beta1, and IL-12p40 levels in Crohn's disease and ulcerative colitis patients in remission versus active disease. *American Journal Gastroenterology, 100*(2), 414–423.

Kaukola, T., Satyaraj, E., Patel, D. D., Tchernev, V. T., Grimwade, B. G., Kingsmore, S. F., Koskela, P., Tammela, O., Vainionpaa, L., Pihko, H., Aarimaa, T., & Hallman, M. (2004). Cerebral palsy is characterized by protein mediators in cord serum. *Annual Neurology, 55*(2), 186–194.

Kingsmore, S. F. (2006). Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nature Reviews, 5*(4), 310–320.

Knezevic, V., Leethanakul, C., Bichsel, V. E., Worth, J. M., Prabhu, V. V., Gutkind, J. S., Liotta, L. A., Munson, P. J., Petricoin Iii, E. F., & Krizman, D. B. (2001). Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics, 1*(10), 1271–1278.

Ko, I. K., Kato, K., & Iwata, H. (2005). Parallel analysis of multiple surface markers expressed on rat neural stem cells using antibody microarrays. *Biomaterials, 26*(23), 4882–4891.

Kopf, E., Shnitzer, D., & Zharhary, D. (2005). Panorama Ab Microarray Cell Signaling kit: A unique tool for protein expression analysis. *Proteomics, 5*(9), 2412–2416.

Krishnan, C., Kaplin, A. I., Graber, J. S., Darman, J. S., & Kerr, D. A. (2005). Recurrent transverse myelitis following neurobrucellosis: Immunologic features and beneficial response to immunosuppression. *Journal of Neurovirology, 11*(2), 225–231.

Kristensson, M., Olsson, K., Carlson, J., Wullt, B., Sturfelt, G., Borrebaeck, C. A. K., & Wingren, C. (2012). Design of recombinant antibody microarrays for urinary proteomics. *Proteomics Clinical Applications, 6*(5–6), 291–296.

Kusnezow, W., Banzon, V., Schroder, C., Schaal, R., Hoheisel, J. D., Ruffer, S., Luft, P., Duschl, A., & Syagailo, Y. V. (2007). Antibody microarray-based profiling of complex specimens: Systematic evaluation of labeling strategies. *Proteomics, 7*(11), 1786–1799.

Lao, Y. H., Peck, K., & Chen, L. C. (2009). Enhancement of aptamer microarray sensitivity through spacer optimization and avidity effect. *Analytical Chemistry, 81*(5), 1747–1754.

Ligler, F. S., Taitt, C. R., Shriver-Lake, L. C., Sapsford, K. E., Shubin, Y., & Golden, J. P. (2003). Array biosensor for detection of toxins. *Analytical Bioanalytical Chemistry, 377*(3), 469–477.

Lin, Y., Huang, R., Chen, L., Li, S., Shi, Q., Jordan, C., & Huang, R. P. (2004). Identification of interleukin-8 as estrogen receptor-regulated factor involved in breast cancer invasion and angiogenesis by protein arrays. *International Journal of Cancer, 109*(4), 507–515.

Lin, M. W., Ho, J. W., Harrison, L. C., dos Remedios, C. G., & Adelstein, S. (2013). An antibody-based leukocyte-capture microarray for the diagnosis of systemic lupus erythematosus. *PLoS One, 8*(3), e58199.

Liu, T., Xue, R., Dong, L., Wu, H., Zhang, D., & Shen, X. (2011). Rapid determination of serological cytokine biomarkers for hepatitis B virus-related hepatocellular carcinoma using antibody microarrays. *Acta Biochimica et Biophysica Sinica Shanghai, 43*(1), 45–51.

Lundberg, K., Lindstedt, M., Larsson, K., Dexlin, L., Wingren, C., Ohlin, M., Greiff, L., & Borrebaeck, C. A. K. (2008). Augmented Phl p 5-specific Th2 response after exposure of dendritic cells to allergen in complex with specific IgE compared to IgG1 and IgG4. *Clinical Immunology (Orlando Fla), 128*(3), 358–365.

Madan, M., Bishayi, B., Hoge, M., Messas, E., & Amar, S. (2007). Doxycycline affects diet- and bacteria-associated atherosclerosis in an ApoE heterozygote murine model: Cytokine profiling implications. *Atherosclerosis, 190*(1), 62–72.

Miller, J. C., Zhou, H., Kwekel, J., Cavallo, R., Burke, J., Butler, E. B., Teh, B. S., & Haab, B. B. (2003). Antibody microarray profiling of human prostate cancer sera: Antibody screening and identification of potential biomarkers. *Proteomics, 3*(1), 56–63.

Mitchell, A. M., Brown, M. D., Menown, I. B., & Kline, J. A. (2005). Novel protein markers of acute coronary syndrome complications in low-risk outpatients: A systematic review of potential use in the emergency department. *Clinical Chemistry, 51*(11), 2005–2012.

Mor, G., Visintin, I., Lai, Y., Zhao, H., Schwartz, P., Rutherford, T., Yue, L., Bray-Ward, P., & Ward, D. C. (2005). Serum protein markers for early detection of ovarian cancer. *Procedings of National Academy Science U S A, 102*(21), 7677–7682.

Nishizuka, S. S., & Mills, G. B. (2016). New era of integrated cancer biomarker discovery using reverse-phase protein arrays. *Drug Metabbolism and Pharmacokinetics, 31*(1), 35–45.

Olsson, N., Wingren, C., Mattsson, M., James, P., O'Connell, D., Nilsson, F., Cahill, D. J., & Borrebaeck, C. A. K. (2011). Proteomic analysis and discovery using affinity proteomics and mass spectrometry. *Molecular & Cellular Proteomics, 10*(10), M110 003962.

Olsson, N., James, P., Borrebaeck, C. A. K., & Wingren, C. (2012a). Quantitative proteomics targeting classes of motif-containing peptides using immunoaffinity-based mass spectrometry. *Molecular & Cellular Proteomics, 11*(8), 342–354.

Olsson, N., Wallin, S., James, P., Borrebaeck, C. A. K., & Wingren, C. (2012b). Epitope-specificity of recombinant antibodies reveals promiscuous peptide-binding properties. *Protein Science, 21*(12), 1897–1910.

Orchekowski, R., Hamelinck, D., Li, L., Gliwa, E., van Brocklin, M., Marrero, J. A., Vande Woude, G. F., Feng, Z., Brand, R., & Haab, B. B. (2005). Antibody microarray profiling reveals individual and combined serum proteins associated with pancreatic cancer. *Cancer Research, 65*(23), 11193–11202.

Parker, C. E., & Borchers, C. H. (2014). Mass spectrometry based biomarker discovery, verification, and validation–quality assurance and control of protein biomarker assays. *Molecular Oncology, 8*(4), 840–858.

Patel, H., Nteliopoulos, G., Nikolakopoulou, Z., Jackson, A., & Gordon, M. Y. (2011). Antibody arrays identify protein-protein interactions in chronic myeloid leukaemia. *Brittish Journal of Haematology, 152*(5), 611–614.

Poetz, O., Hoeppe, S., Templin, M. F., Stoll, D., & Joos, T. O. (2009). Proteome wide screening using peptide affinity capture. *Proteomics, 9*(6), 1518–1523.

Rahman, A., & Isenberg, D. A. (2008). Systemic lupus erythematosus. *New England Journal of Medicine, 358*(9), 929–939.

Ramachandran, N., Hainsworth, E., Bhullar, B., Eisenstein, S., Rosen, B., Lau, A. Y., Walter, J. C., & LaBaer, J. (2004). Self-assembling protein microarrays. *Science, 305*(5680), 86–90.

Ramachandran, N., Hainsworth, E., Demirkan, G., & LaBaer, J. (2006). On-chip protein synthesis for making microarrays. *Methods in Molecular Biology (Clifton, NJ), 328*, 1–14.

Ramachandran, N., Raphael, J. V., Hainsworth, E., Demirkan, G., Fuentes, M. G., Rolfs, A., Hu, Y., & LaBaer, J. (2008). Next-generation high-density self-assembling functional protein arrays. *Nature Methods, 5*(6), 535–538.

Ramirez, A. B., & Lampe, P. D. (2010). Discovery and validation of ovarian cancer biomarkers utilizing high density antibody microarrays. *Cancer Biomarker, 8*(4–5), 293–307.

Renberg, B., Shiroyama, I., Engfeldt, T., Nygren, P. K., & Karlstrom, A. E. (2005). Affibody protein capture microarrays: Synthesis and evaluation of random and directed immobilization of affibody molecules. *Analytical Biochemistry, 341*(2), 334–343.

Renberg, B., Nordin, J., Merca, A., Uhlen, M., Feldwisch, J., Nygren, P. A., & Karlstrom, A. E. (2007). Affibody molecules in protein capture microarrays: Evaluation of multidomain ligands and different detection formats. *Journal of Proteome Research, 6*(1), 171–179.

Rovin, B. H., & Zhang, X. (2009). Biomarkers for lupus nephritis: The quest continues. *Clinical Journal of the American Society of Nephrology, 4*(11), 1858–1865.

Rowe, C. A., Tender, L. M., Feldstein, M. J., Golden, J. P., Scruggs, S. B., MacCraith, B. D., Cras, J. J., & Ligler, F. S. (1999). Array biosensor for simultaneous identification of bacterial, viral, and protein analytes. *Analytical Chemistry, 71*(17), 3846–3852.

Rowe-Taitt, C. A., Golden, J. P., Feldstein, M. J., Cras, J. J., Hoffman, K. E., & Ligler, F. S. (2000). Array biosensor for detection of biohazards. *Biosensor and Bioelectronics, 14*(10–11), 785–794.

Rubina, A. Y., Dyukova, V. I., Dementieva, E. I., Stomakhin, A. A., Nesmeyanov, V. A., Grishin, E. V., & Zasedatelev, A. S. (2005). Quantitative immunoassay of biotoxins on hydrogel-based protein microchips. *Analytical Biochemistry, 340*(2), 317–329.

Rucker, V. C., Havenstrite, K. L., & Herr, A. E. (2005). Antibody microarrays for native toxin detection. *Analytical Biochemistry, 339*(2), 262–270.

Saerens, D., Ghassabeh, G. H., & Muyldermans, S. (2008). Antibody technology in proteomics. *Brief Functional Genomic Proteomic, 7*(4), 275–282.

Sanchez-Carbayo, M., Socci, N. D., Lozano, J. J., Haab, B. B., & Cordon-Cardo, C. (2006). Profiling bladder cancer using targeted antibody arrays. *American Journal of Pathology, 168*(1), 93–103.

Sandstrom, A., Andersson, R., Segersvard, R., Lohr, M., Borrebaeck, C. A. K., & Wingren, C. (2012). Serum proteome profiling of pancreatitis using recombinant antibody microarrays reveals disease-associated biomarker signatures. *Proteomics Clinical Applications, 6*(9–10), 486–496.

Schroder, C., Jacob, A., Tonack, S., Radon, T. P., Sill, M., Zucknick, M., Ruffer, S., Costello, E., Neoptolemos, J. P., Crnogorac-Jurcevic, T., Bauer, A., Fellenberg, K., & Hoheisel, J. D. (2010). Dual-color proteomic profiling of complex samples with a microarray of 810 cancer-related antibodies. *Molecular & Cellular Proteomics, 9*(6), 1271–1280.

Schroder, C., Alhamdani, M. S., Fellenberg, K., Bauer, A., Jacob, A., & Hoheisel, J. D. (2011). Robust protein profiling with complex antibody microarrays in a dual-colour mode. *Methods Molecular Biology, 785*, 203–221.

Schwenk, J. M., Gry, M., Rimini, R., Uhlen, M., & Nilsson, P. (2008). Antibody suspension bead arrays within serum proteomics. *Journal of Proteome Research, 7*(8), 3168–3179.

Shafer, M. W., Mangold, L., Partin, A. W., & Haab, B. B. (2007). Antibody array profiling reveals serum TSP-1 as a marker to distinguish benign from malignant prostatic disease. *Prostate, 67*(3), 255–267.

Sharma, M., Arnason, J. T., Burt, A., & Hudson, J. B. (2006). Echinacea extracts modulate the pattern of chemokine and cytokine secretion in rhinovirus-infected and uninfected epithelial cells. *Phytother Research, 20*(2), 147–152.

Shi, W., Meng, Z., Chen, Z., Luo, J., & Liu, L. (2011). Proteome analysis of human pancreatic cancer cell lines with highly liver metastatic potential by antibody microarray. *Molecular & Cellular Biochemistry, 347*(1–2), 117–125.

Smith, L., Watson, M. B., O'Kane, S. L., Drew, P. J., Lind, M. J., & Cawkwell, L. (2006). The analysis of doxorubicin resistance in human breast cancer cells using antibody microarrays. *Molecular Cancer Therapy, 5*(8), 2115–2120.

Soderlind, E., Strandberg, L., Jirholt, P., Kobayashi, N., Alexeiva, V., Aberg, A. M., Nilsson, A., Jansson, B., Ohlin, M., Wingren, C., Danielsson, L., Carlsson, R., & Borrebaeck, C. A. K. (2000). Recombining germline-derived CDR sequences for creating diverse single-framework antibody libraries. *Nature Biotechnology, 18*(8), 852–856.

Sokolov, B. P., & Cadet, J. L. (2006). Methamphetamine causes alterations in the MAP kinase-related pathways in the brains of mice that display increased aggressiveness. *Neuropsychopharmacology, 31*(5), 956–966.

Solier, C., & Langen, H. (2014). Antibody-based proteomics and biomarker research – current status and limitations. *Proteomics, 14*(6), 774–783. doi:10.1002/pmic.201300334.

Sreekumar, A., Nyati, M. K., Varambally, S., Barrette, T. R., Ghosh, D., Lawrence, T. S., & Chinnaiyan, A. M. (2001). Profiling of cancer cells using protein microarrays: Discovery of novel radiation-regulated proteins. *Cancer Research, 61*(20), 7585–7593.

Srivastava, M., Eidelman, O., Jozwik, C., Paweletz, C., Huang, W., Zeitlin, P. L., & Pollard, H. B. (2006). Serum proteomic signature for cystic fibrosis using an antibody microarray platform. *Molecular Genetics Metabolism, 87*(4), 303–310.

Sukhdeo, K., Paramban, R. I., Vidal, J. G., Elia, J., Martin, J., Rivera, M., Carrasco, D. R., Jarrar, A., Kalady, M. F., Carson, C. T., Balderas, R., Hjelmeland, A. B., Lathia, J. D., & Rich, J. N. (2013). Multiplex flow cytometry barcoding and antibody arrays identify surface antigen profiles of primary and metastatic colon cancer cell lines. *PLoS One, 8*(1), e53015.

Sun, H., Chua, M. S., Yang, D., Tsalenko, A., Peter, B. J., & So, S. (2008). Antibody arrays identify potential diagnostic markers of hepatocellular carcinoma. *Biomarker Insights, 3*, 1–18.

Svedhem, S., Pfeiffer, I., Larsson, C., Wingren, C., Borrebaeck, C. A. K., & Hook, F. (2003). Patterns of DNA-labeled and scFv-antibody-carrying lipid vesicles directed by material-specific immobilization of DNA and supported lipid bilayer formation on an Au/SiO2 template. *Chembiochem, 4*(4), 339–343.

Szodoray, P., Alex, P., Brun, J. G., Centola, M., & Jonsson, R. (2004). Circulating cytokines in primary Sjogren's syndrome determined by a multiplex cytokine array system. *Scandinavian Journal Immunology, 59*(6), 592–599.

Taitt, C. R., Anderson, G. P., Lingerfelt, B. M., Feldstein, S. M., & Ligler, F. S. (2002). Nine-analyte detection using an array-based biosensor. *Analytical Chemistry, 74*(23), 6114–6120.

Tannapfel, A., Anhalt, K., Hausermann, P., Sommerer, F., Benicke, M., Uhlmann, D., Witzigmann, H., Hauss, J., & Wittekind, C. (2003). Identification of novel proteins associated with hepatocellular carcinomas using protein microarrays. *Journal of Pathology, 201*(2), 238–249.

Tuomisto, T. T., Riekkinen, M. S., Viita, H., Levonen, A. L., & Yla-Herttuala, S. (2005). Analysis of gene and protein expression during monocyte-macrophage differentiation and cholesterol loading–cDNA and protein array study. *Atherosclerosis, 180*(2), 283–291.

Turtinen, L. W., Prall, D. N., Bremer, L. A., Nauss, R. E., & Hartsel, S. C. (2004). Antibody array-generated profiles of cytokine release from THP-1 leukemic monocytes exposed to different amphotericin B formulations. *Antimicrobial Agents and Chemotherapy, 48*(2), 396–403.

Uhlen, M., & Hober, S. (2009). Generation and validation of affinity reagents on a proteome-wide level. *Journal Molecular Recognition, 22*(2), 57–64.

Uhlen, M., & Ponten, F. (2005). Antibody-based proteomics for human tissue profiling. *Molecular & Cellular Proteomics, 4*(4), 384–393.

Vazquez-Martin, A., Colomer, R., & Menendez, J. A. (2007). Protein array technology to detect HER2 (erbB-2)-induced 'cytokine signature' in breast cancer. *European Journal of Cancer, 43*(7), 1117–1124.

Volk, S., Schreiber, T. D., Eisen, D., Wiese, C., Planatscher, H., Pynn, C. J., Stoll, D., Templin, M. F., Joos, T. O., & Potz, O. (2012). Combining ultracentrifugation and peptide termini group-specific immunoprecipitation for multiplex plasma protein analysis. *Molecular & Cellular Proteomics, 11*(7), O111 015438.

Voshol, H., Ehrat, M., Traenkle, J., Bertrand, E., & van Oostrum, J. (2009). Antibody-based proteomics: Analysis of signaling networks using reverse protein arrays. *FEBS Journal, 276*(23), 6871–6879.

Wacker, R., & Niemeyer, C. M. (2004). DDI-microFIA–A readily configurable microarray-fluorescence immunoassay based on DNA-directed immobilization of proteins. *Chembiochem, 5*(4), 453–459.

Wacker, R., Schroder, H., & Niemeyer, C. M. (2004). Performance of antibody microarrays fabricated by either DNA-directed immobilization, direct spotting, or streptavidin-biotin attachment: A comparative study. *Analytical Biochemisty, 330*(2), 281–287.

Walter, J. G., Kokpinar, O., Friehs, K., Stahl, F., & Scheper, T. (2008). Systematic investigation of optimal aptamer immobilization for protein-microarray applications. *Analytical Chemistry, 80*(19), 7372–7378.

Wang, C. C., Yim, K. W., Poon, T. C., Choy, K. W., Chu, C. Y., Lui, W. T., Lau, T. K., Rogers, M. S., & Leung, T. N. (2007). Innate immune response by ficolin binding in apoptotic placenta is associated with the clinical syndrome of preeclampsia. *Clinical Chemistry, 53*(1), 42–52.

Whiteaker, J. R., Zhao, L., Zhang, H. Y., Feng, L. C., Piening, B. D., Anderson, L., & Paulovich, A. G. (2007). Antibody-based enrichment of peptides on magnetic beads for mass-spectrometry-based quantification of serum biomarkers. *Analytical Biochemistry, 362*(1), 44–54.

Whiteaker, J. R., Zhao, L., Anderson, L., & Paulovich, A. G. (2010). An automated and multiplexed method for high throughput peptide immunoaffinity enrichment and multiple reaction monitoring mass spectrometry-based quantification of protein biomarkers. *Molecular & Cellular Proteomics, 9*(1), 184–196.

Wingren, C., & Borrebaeck, C. A. K. (2004). High-throughput proteomics using antibody microarrays. *Expert Review of Proteomics, 1*(3), 355–364.

Wingren, C., & Borrebaeck, C. A. K. (2006). Antibody microarrays – Current status and key technological advances. *OMICS, 10*(3), 411–427.

Wingren, C., & Borrebaeck, C. A. K. (2007). Progress in miniaturization of protein arrays–a step closer to high-density nanoarrays. *Drug Discovery Today, 12*(19–20), 813–819.

Wingren, C., & Borrebaeck, C. A. K. (2008). Antibody microarray analysis of directly labelled complex proteomes. *Current Opinion in Biotechnology, 19*(1), 55–61.

Wingren, C., & Borrebaeck, C. A. K. (2009). Antibody-based microarrays. *Methods in Molecular Biology (Clifton, NJ), 509*, 57–84.

Wingren, C., Ingvarsson, J., Dexlin, L., Szul, D., & Borrebaeck, C. A. K. (2007). Design of recombinant antibody microarrays for complex proteome analysis: Choice of sample labeling-tag and solid support. *Proteomics, 7*(17), 3055–3065.

Wingren, C., James, P., & Borrebaeck, C. A. K. (2009). Strategy for surveying the proteome using affinity proteomics and mass spectrometry. *Proteomics, 9*(6), 1511–1517.

Wingren, C., Sandström, A., Segersvärd, R., Carlsson, A., Andersson, R., Löhr, M., & Borrebaeck, C. A. K. (2012). Identification of serum biomarker signatures associated with pancreatic cancer. *Cancer Research, 72*(10), 2481–2490.

Wong, J., Sibani, S., Lokko, N. N., LaBaer, J., & Anderson, K. S. (2009). Rapid detection of antibodies in sera using multiplexed self-assembling bead arrays. *Journal of Immunological Methods, 350*(1–2), 171–182.

Wu, A. H., Smith, A., Christenson, R. H., Murakami, M. M., & Apple, F. S. (2004). Evaluation of a point-of-care assay for cardiac markers for patients suspected of acute myocardial infarction. *Clinica Chimica Acta, 346*(2), 211–219.

Yee, S. B., Bourdi, M., Masson, M. J., & Pohl, L. R. (2007). Hepatoprotective role of endogenous Interleukin-13 in a Murine Model of Acetaminophen-Induced Liver Disease. Chemical Research in Toxicology

Yue, T., Goldstein, I. J., Hollingsworth, M. A., Kaul, K., Brand, R. E., & Haab, B. B. (2009). The prevalence and nature of glycan alterations on specific proteins in pancreatic cancer patients revealed using antibody-lectin sandwich arrays. *Molecular and Cellular Proteomics, 8*(7), 1697–1707.

Yue, T., Partyka, K., Maupin, K. A., Hurley, M., Andrews, P., Kaul, K., Moser, A. J., Zeh, H., Brand, R. E., & Haab, B. B. (2011). Identification of blood-protein carriers of the CA 19–9 antigen and characterization of prevalence in pancreatic diseases. *Proteomics, 11*(18), 3665–3674.

Zhang, Y., Lou, J., Jenko, K. L., Marks, J. D., & Varnum, S. M. (2012). Simultaneous and sensitive detection of six serotypes of botulinum neurotoxin using enzyme-linked immunosorbent assay-based protein antibody microarrays. *Analytical Biochemistry, 430*(2), 185–192.

Zhou, H., Bouwman, K., Schotanus, M., Verweij, C., Marrero, J. A., Dillon, D., Costa, J., Lizardi, P., & Haab, B. B. (2004). Two-color, rolling-circle amplification on antibody microarrays for sensitive, multiplexed serum-protein measurements. *Genome Biology, 5*(4), R28.

Zhou, Q., Desta, T., Fenton, M., Graves, D. T., & Amar, S. (2005). Cytokine profiling of macrophages exposed to Porphyromonas gingivalis, its lipopolysaccharide, or its FimA protein. *Infectious Immunity, 73*(2), 935–943.

Index

Á. Végvári (ed.), *Proteogenomics*, Advances in Experimental Medicine and Biology 926,
DOI 10.1007/978-3-319-42316-6