8

Computational Approaches in Proteomics

Karla Cervantes Gracia[1] • Holger Husi[2,3]

[1]Basic Sciences Division, Universidad de Monterrey, San Pedro Garza García, N.L. Mexico; [2]Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK; [3]Division of Biomedical Sciences, Centre for Health Science, University of Highlands and Islands, Inverness, UK

Author for correspondence: Holger Husi, Division of Biomedical Sciences, University of the Highlands and Islands, Centre for Health Science, Inverness IV2 3JH, UK.
Email: Holger.Husi@uhi.ac.uk

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch8

Abstract: Understanding of biological processes and aberrations in disease conditions has over the years moved away from the study of single molecules to a more holistic and all-encompassing view to investigate the entire spectrum of proteins. This method, termed proteomics, has been enabled principally by mass spectrometry techniques. The power of mass spectrometry-based proteomics lays in its ability to investigate an entire proteome and associated expression or modification states of a huge amount of proteins in one single experiment. This massive amount of data requires a high level of automation in data processing to render it into a reduced set of information that can be used to answer the initial hypotheses, explore the biology or contextualize molecular changes associated with a physiological attribute. This chapter gives an overview of the most common proteomic approaches, biological sample considerations and data acquisition methods, data processing, software solutions for the various steps and further functional analyses of biological data. This enables the comparison of various datasets as a summation of individual experiments, to cross-compare sample types and other metadata. There are many approach pipelines in existence that cover specialist disciplines and data analytics steps, and it is a certainty that many more data analysis methodologies will be generated over the coming years, but it also emphasizes the

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia.
ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

inherent place of proteomic technologies in research in elucidating the nature of biological processes and understanding of disease etiology.

Keywords: data analysis; mass spectrometry; proteomics; software; workflows

INTRODUCTION

The development and improvement of high-throughput techniques in "omic" science have paved the way not only to a broader view of the molecules involved in a specific condition but also to generate networks of all interacting elements (genes, proteins, and metabolites) to gain a better understanding of how a specific biological system works. Despite the over-abundance of genomics research in this field, there is so much more complexity left out in a system that can be explained by the understanding and integration of proteomic data. The proteome is more complex and is not as stable as the genome, and it is not only based on what is observed in the genome but also influenced by several factors. Protein expression depends on tissue type, environmental stimuli, and post-translational modifications (PTM) that influence its level of activity, structure, function and regulation (1, 2). Moreover, life depends on proteins, as they are responsible for many complex processes within a cell, from replication, gene transcription and translation to cellular senescence and death. Therefore, by having a better understanding of the proteome, a wider comprehension of cellular regulation can be achieved. Proteomics is the high-throughput study of proteins incorporating the identification, quantitation, analysis and comparison of differential expression of proteins from samples under specific biological conditions. The characterization of the proteome involves the identification of structure, function, interactions and modifications (3).

Because of its improved sensitivity and specificity, mass spectrometry (MS) proteomics is the most widely used approach, and it is considered the method of choice to obtain global measurements of proteins (4). The most common and classic applications of proteomics are to characterize large datasets to create an inventory of identified proteins in different tissue or cellular samples, as well as to generate lists of differentially expressed proteins from samples under specific conditions (5). However, these data alone lack a biological meaning, and therefore, it is essential to pursue additional approaches to allow a better interpretation of biological processes (6, 7).

Qualitative and quantitative methods are also of importance in network analyses. Qualitative approaches are much more common. Although quantitative network analysis can generate more specialized results and are better adapted to generate new insights and advancement in biomedical research by unraveling the significant proteins that interplay in a disease, producing new diagnostic hypothesis, the standardization and homogenization of its analysis still need improvement to establish its reliability and reproducibility when analyzing high-throughput data (8, 9).

High-throughput technologies and bioinformatic tools are fundamental for proteomics data interpretation to discover new biological insights on cellular processes, disease etiology and biomarker candidates. Although these tools are under

continuous update and new approaches are implemented, the development of harmonized benchmarks for datasets and analysis, as well as to establish gold-standard workflows, is imperative to produce more reliable and reproducible results, and by doing so, it will help to overcome the challenges of proteomics data interpretation (10, 11).

The acquisition of a vast amount of high-quality raw spectra using MS is nowadays a relatively simple task with the right equipment and involves a high level of automation, which however leads to a fundamental, and crucial, step to mathematically and statistically interrogate the data and ultimately match it to a library of known or hypothetical molecules. This is of particular importance in strategies such as shotgun proteomics and other large-scale MS screens, whereas specific applications such as selective or multiple reaction monitoring (SRM/MRM) have a different requirement for the entire workflow and require appropriate specific software solutions (12). Figure 1 shows a general overview of a typical proteomics workflow, starting from protein and peptide preparation from tissues to the computational procedures to obtain a list of molecules with associated confidence or significance scores that can then be analyzed further.

SAMPLE TYPES AND SAMPLE PROCESSING APPROACHES

Body homeostasis is maintained through specialized systems, which are orchestrated by the interplay between cells, tissues and organs. Each system anomaly can be better described by specific samples; therefore, to characterize diseases, it becomes essential to analyze the proteome of the appropriate samples. Many sample types are suitable for proteomic analysis, including cells, organs, tissues and body fluids. Biomarker discovery helps to identify pathological states, track disease progression and improve diagnostics or disease etiology, which are some of the common applications when using these sample types (13). An important

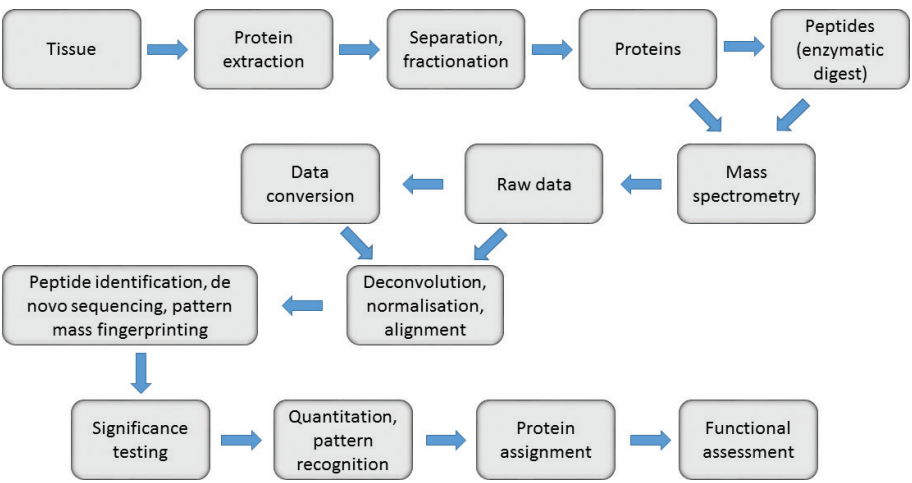

Figure 1 Flowchart and procedures for a generic proteomics pipeline.

factor to be considered for the success of a proteomic approach is the quality and quantity of the sample, due to the challenge that its complexity implies for MS techniques. As the detection rates of proteins using MS are directly related to the absolute quantity of these biomolecules in a sample, high-abundance proteins tolerate losses during processing quite well. However, the detection rates of low-abundance proteins are usually much more sensitive towards loss due to common instrumentation and processes, and therefore, the preferred workflow is microproteomics, which minimizes this loss and increases the sample processing efficiency (14). Samples, such as the ones derived from cancer or tumour cells, exemplify low-abundance protein samples. They should be analyzed by specific microproteomic workflows with special adaptations of the techniques for sample preparation, cleaning, fractionation and separation to ensure minimal losses before analysis and increase sensitivity of nano/microgram-samples that allows maximal identification of low-abundance proteins (15, 16).

Cell lines

A general overview of disadvantages and applications of each sample-source type is presented in Table 1. Although heterogeneous cell populations that compose tissues can be individually isolated and analyzed, cell lines are believed to reflect the protein composition of primary cells and specific tissues. Moreover, the reproducibility in proteomic analysis using cell lines is one of its main advantages over other sample types. It also allows proteomic subcellular analysis (17). Several applications of proteomic analyses using cell lines have been established to investigate molecular pathways of specified cell types, differences between normal and disease phenotypes, and different stages of diseases (18, 19). However, problems in cell line culturing are rather common if no quality control is carried out and can lead to unreliable results if not detected and treated. The most common problems with cell lines are genomic instability, infections by microorganisms that could alter cell turnover and protein expression patterns and cross-contamination leading to the growth of a mixture of cell types, affecting the results of the study even before proteomic analysis can be performed (20).

Tissue culture

Proteomic analysis with tissue culture as a sample is also a very informative approach. It allows the interaction and analysis of the diverse cell types involved in a disease, leading to a broader view of the biological systems of importance in pathology. It is basically based on the growth of tissue outside the organism, under controlled conditions. Tissue samples for this are obtained through surgery from humans or animals. Tissue culture-based proteomic profiling is useful for understanding the biological mechanisms underlying a disease, biomarker and therapeutic target identification as well as effects in a sample due to viral, drug or genetic changes (21). Techniques, such as 3D co-culture systems, fresh tissue proteomics and tumour spheroid models, have improved the analysis and results (22). However, its accessibility remains as its major downside, and no accurate track in disease progression can be performed without re-sampling.

Organ

Organ samples can be maintained under specific culture conditions, and its different cell types can be analyzed. It is the most difficult sample to obtain from humans, and since biofluids are secreted from several organs and make proteins more accessible, they are the sample of choice for biomarker discovery and pathology research (23, 24). However, reliability and reproducibility are still issues to be addressed, before they can be eventually established as a good source of clinical proteomics. Like the tissue samples, animal models serve as a good source of organ samples. They provide a controlled environment and the possibility to follow up the changes in proteomic profiling throughout the course of a disease. The major drawback using animal models is that they cannot accurately predict how a system works in humans (25). However, in order to overcome this issue and to have a better and broader understanding of the interaction of human proteome within a system, new engineered model systems have been created, such as multiorgan lab-on-a-chip platforms, that show a better correlation with human systems than animal models, mimicking the key aspects of responses like drug treatment (26, 27).

Exosomes

Besides the analysis of the proteome in cells, proteins secreted by the cells have gained attention when unraveling the etiology of diseases. All together, these proteins are known as secretome, and a specific component of the secretome that has been studied in relation to pathology is the exosome. Exosomes are membrane vesicles, differentiated from other vesicles by size and expression of the CD81 protein. They have a very low abundance of proteins, which is undetectable using biofluid analysis (28). Among these proteins are some that are specific to the biological fluid or cell, making the exosomes an interesting source for biomarkers to advance the identification and understanding of pathologies (29).

Biological fluids

Depending on the purpose of the research, diverse body fluids can be collected and processed for proteomic analysis. A general overview of disadvantages and applications of each sample-source type is presented in Table 1. Among the commonly analyzed biological fluids in proteomics are blood, serum, plasma, cerebrospinal fluid (CSF), urine, saliva and semen. The fluctuation in their protein levels is expected to reflect pathophysiological conditions; however, some drawbacks such as protein content, high abundance of masking proteins, and sample instability can lead to complex interpretations (30). Blood, serum and plasma are the most common biological fluids in proteomic research due to its non-invasive nature and its high concentration of protein/peptides, as well as the assumption that blood reflects the pathophysiological state of several organs. Biofluids such as urine and CSF are not the most desirable samples for proteomics because they contain a lower protein/peptide concentration (31). In addition, a complicated collection process is a hindrance in obtaining a reasonable amount of CSF sample (31).

TABLE 1 **Biological fluids overview: Applications and disadvantages**

Biological fluid	Applications	Disadvantages
Serum and plasma	Serum and plasma have been used for multiple proteomics–based biomarker discovery studies.	Dynamic qualitative and quantitative range of proteins; small number of highly abundant proteins can mask potential biomarkers; biomarker of interest can be lost upon the removal of highly abundant proteins.
Cerebrospinal fluid (CSF)	Potential diagnostic utility in neurodegenerative diseases including Alzheimer's, multiple sclerosis and Parkinson's.	Requires a lumbar puncture or a spinal tap, invasive procedures. Traumatic punctures can alter CSF protein expression levels and skew a diagnosis; small volumes of samples obtained; yield a highly dynamic range of protein concentrations; small number of highly abundant proteins can mask potential biomarkers; depletion techniques are neither time nor cost-effective techniques; biomarker of interest can be lost upon the removal of highly abundant proteins.
Urine	Good source of biomarkers for urogenital and systemic diseases.	Definition of disease-specific biomarkers is complicated; significant changes in the proteome throughout the day can be connected with the time of collection, fluid intake, diet, exercise, circadian rhythms and circulatory levels of various hormones; presence of MS hampering salts; lower concentration of proteins/peptides compared to serum and plasma.
Saliva	Most of the biomolecules that are usually detected in urine and blood can also be found in salivary secretions; about 30% of blood proteins are also present in saliva.	Very low concentration of proteins; very rapid protein degradation in whole saliva at room temperature, this may occur during saliva collection and handling.
Semen	Applications in research areas such as reproduction and prostate cancer, and used for many purposes in the diagnosis of male fertility.	Small number of highly abundant proteins can mask potential biomarkers; biomarker of interest can be lost upon the removal of high abundant proteins.
Circulating tumour cells (CTC)	Practical application in diagnosis and disease treatment, determine the prognosis of metastatic progression or relapse, monitor anti-cancer treatments, understand the mechanism of metastatic disease and develop new strategies in disease treatment.	Very low abundance of CTC in blood; cell heterogeneity makes it difficult to isolate the whole CTC population.

Although the technicalities of sample collection, management and storage are known to be of vital importance to keep the composition and quality of the sample to be reproducible and reliable, there is no commonly accepted standardization protocol for bio-sampling procedures. Variables, such as storage times; temperatures and number of freeze-thaw cycles; removal of additives, such as heparin to prevent clotting; as well as the consumables, such as collection and processing tubes, are important parameters to be considered in order to avoid differences in protein composition among samples (31, 32). Bio-sampling optimization and standardization are essential steps to improve reproducibility for accurate correlations among different studies (33).

DATA ACQUISITION

Proteomics has become a feasible and a promising approach with the advancements in MS methodologies. MS/MS innovations and possible combinations are constantly under improvement, and nowadays, it has become the gold standard for any kind of proteomic studies. Furthermore, high-resolution mass spectrometers have been recently adapted for high-throughput proteomics (34). MS/MS has a high impact in lowering sample complexity by the isolation of precursor ions through a mass filter, as well as their fragmentation and further detection by high-resolution mass analyzers (35). Moreover, for each of these steps, technologies have been developed to identify and distinguish peptides more accurately, with a better resolution, coverage and reproducibility. Also, computer tools have been under constant development to improve the analysis of the complex outcome data (Table 2). In order to achieve a more accurate protein identification, three main approaches have been described: bottom–up (BU), top–down (TD) and, more recently, middle–down (MD) (Figure 2).

Bottom–up data analysis

In contrast to TD and MD proteomics analysis, for BU data analysis, a deconvolution step is not required when implementing ESI, due to the rare generation of double- and triple-charged fragment ions (36). Mass spectra raw data are commonly processed by Proteome Discoverer or MaxQuant platforms using several search engines, such as Sequest, Mascot, Andromeda, X!Tandem and COMET, usually against UniProt databases (37–39). MaxQuant software can also determine protein quantitation and estimate the error of PTM false localization. For downstream correlation and clustering analysis, the identified proteins are commonly processed in the Perseus platform (38, 40, 41). To reduce data complexity, principal component analysis (PCA) has been the method of choice, and also to identify the relatedness of the differentially expressed proteins within and among samples (39, 41). Moreover, to interpret the potential function of the datasets obtained, the DAVID platform is commonly used to enrich them with Gene Ontology terms, KEGG pathway information and InterPro protein domains (39, 42). Additionally, constructed networks are commonly visualized in Cytoscape, and in order to identify functional and physical associations among mRNA and protein data, the STRING database is used (43). All MS data are usually deposited

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses**

Name	Scope	URL
Reference databases		
UniProt/SwissProt	Proteome assemblies	https://www.uniprot.org
RefSEQ	Genome, transcriptome and proteome assemblies	https://www.ncbi.nlm.nih.gov/refseq/
Tandem mass spectra protein/peptide search engines		
SEQUEST, Comet	Cross-correlation-based scoring, commercial (SEQUEST), open source, free (Comet)	https://proteomicsresource.washington.edu/protocols06/sequest.php
Mascot	Probability-based scoring, commercial	http://www.matrixscience.com/
X!Tandem	Statistical confidence (expectation value) scoring, pattern matching, open source	https://www.thegpm.org/tandem/
Andromeda	Probabilistic scoring model, open source	http://coxdocs.org/doku.php?id=maxquant:andromeda:start
ProLuCID	Three-tier scoring system, binomial probability, cross-correlation calculated Z-score, open source	http://fields.scripps.edu/yates/wp/?page_id=17
Platforms, integrated pipelines		
Proteome Discoverer	Commercial, fully integrated solution for Thermo Fisher instruments, outputs protein/peptide lists, also works with other data formats	https://planetorbitrap.com/proteome-discoverer
Progenesis QI	Commercial, integrated solution from Waters, outputs protein/peptide lists, also works with other data formats	http://www.nonlinear.com/progenesis/qi-for-proteomics/
ProteinPilot	Commercial, fully integrated solution for AB Sciex instruments, outputs protein/peptide lists, also works with other data formats	https://sciex.com/products/software/proteinpilot-software
Integrated Proteomics Pipeline (IP2)	Commercial, uses the ProLuCID search engine, outputs protein/peptide lists, offers limited downstream analysis	http://www.integratedproteomics.com/products.html
Trans-Proteomic Pipeline (TPP)	Modular open-source standardized data processing pipeline	http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
PatternLab	Open-source raw MS file data processing, includes the Comet peptide search engine, Microsoft Windows specific	http://patternlabforproteomics.org/index.html
MaxQuant	Open-source data processing workflow, includes Anromeda peptide search engine, Linux and Windows distributions	https://www.maxquant.org/
Skyline	Modular open-source standardized data processing pipeline, workflow editor, Microsoft Windows specific	https://skyline.ms/project/home/software/Skyline/begin.view
OpenMS	Modular open-source standardized data processing pipeline, workflow editor, OS system independent	https://www.openms.de/
Taverna	Framework for biocomputational tools, workflow editor, open source, re-use of existing workflows, Java programming language based, web server application	https://taverna.incubator.apache.org/
Galaxy	Framework for biocomputational tools, workflow editor, open source, web server application	https://usegalaxy.org/
Data repositories		
PRoteomicsIDEntification database (PRIDE)	Protein and peptide identifications, post-translational modifications, raw data	http://www.ebi.ac.uk/pride/
PeptideAtlas	Library of identified peptides	http://www.peptideatlas.org
Japan ProteOmeSTandard Repository/Database (jPOST)	Integrated proteome datasets	https://jpostdb.org/
MassIVE	Protein and peptide identifications	https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp
iProX	Proteomic datasets	https://www.iprox.org/

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
Panorama Public	Skyline processed data	https://panoramaweb.org/project/Panorama%20Public/begin.view?
Open Proteomics Database (OPD)	Proteomic datasets	http://data.marcottelab.org/MSdata/OPD/
The Global Proteome Machine (GPM)	Metadata	http://gpmdb.thegpm.org/
Functional analysis resources		
Bioconductor/R	Statistical and graphical environment for analysis of high-throughput data	https://www.bioconductor.org/
MixOmics	Statistical package, visualization of analysis runs, requires the statistical software R	http://mixomics.org/
PANDA-view	Statistical analysis, data visualization, quantitative proteomics, requires the statistical software R but provides its own GUI, Microsoft Windows application	https://sourceforge.net/projects/panda-view/
Perseus	Post-analysis of MaxQuant data, data integration, statistical analysis, sample comparisons, Microsoft Windows application	http://coxdocs.org/doku.php?id=perseus:start
InterPro	Protein families, domains and functional sites,	http://www.ebi.ac.uk/interpro/
Gene Ontology (GO)	Hierarchically clustered annotations of functional terms that describe the biological process, molecular function or cellular component	http://www.geneontology.org
Database for Annotation, Visualisation, and Integrated Discovery (DAVID)	GO analysis, KEGG mapping, domain grouping, web application	https://david.ncifcrf.gov/

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
ClueGO	GO analysis, statistical analysis, hierarchical clustering, Cytoscape app	http://www.ici.upmc.fr/cluego/
ReviGO	GO term clustering, semantic analysis	http://revigo.irb.hr/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Pathway mapping, static maps	http://www.genome.jp/kegg/
ReactomeKnowledgeBase	Pathway mapping	http://www.reactome.org
Ingenuity Pathway Knowledge Base (IPA)	Pathway mapping, data clustering, enrichment analysis	https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/
Wikipathways	Database of pathway maps	http://wikipathways.org/index.php/WikiPathways
BioCyc Knowledge Library	Database, pathway mapping	http://biocyc.org/
RHEA	Biochemical reactions database	http://www.rhea-db.org
Protein ANalysisTHrough Evolutionary Relationships (PANTHER)	Signaling pathways	http://www.pantherdb.org
PathVisio	Pathway drawing and pathway analysis tool	https://www.pathvisio.org/
IMPaLA	Pathway analysis, web application	http://impala.molgen.mpg.de/
Molecular Interaction Database (IntAct)	Database, protein-protein interactions, protein-compound interactions	http://www.ebi.ac.uk/intact
Molecular Interaction Database (MINT)	Protein–protein interactions	http://mint.bio.uniroma2.it/

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	Protein–protein interactions	http://string-db.org
JasparDB	Transcription factors and regulatory sites	http://api.bioinfo.no/wsdl/JasparDB.wsdl
Online Mendelian Inheritance in Man (OMIM)	Online catalogue of human genes and genetic disorders	https://www.omim.org/
DisGeNET	Human gene-disease associations, animal models, database, web-query, Ctoscape implementation	http://www.disgenet.org/
Babelomics	Correlations, clustering, ontologies, pathways, heatmaps	http://www.babelomics.org/
Cytoscape	Network analysis environment, graph drawing	https://cytoscape.org/

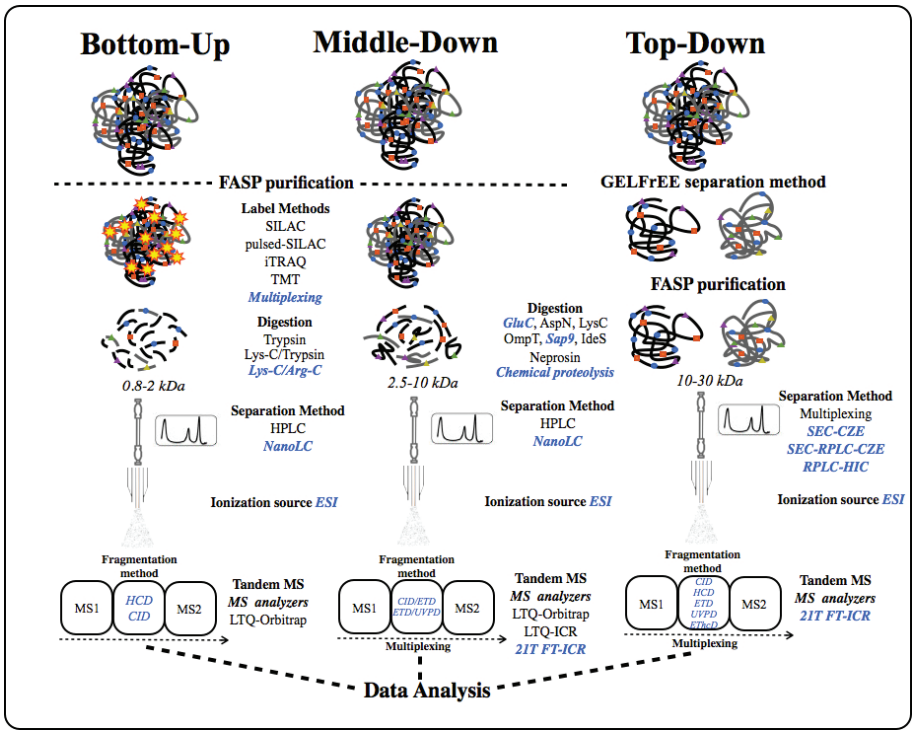

Figure 2 Bottom–up, middle–down and top–down proteomic high-throughput approaches. A general view of each of the approaches and the essential steps to follow are shown from top to bottom. Up-to-date tools, methodologies and techniques most commonly and successfully applied for high-throughput proteomic analyses are highlighted in blue for each of the approaches.

in the ProteomeXchange Consortium via the PRIDE partner repository for sharing, general availability and further study (44, 45).

Top–down data analysis

In TD data analysis, Proteome Discoverer is commonly used to process raw data files, and through its ProSight tools, as well as through MascotTD, identification and characterization of intact proteins can be achieved (46, 47). A database search using ProSight against specific databases (UniProt, SwissProt and RefSEQ) leads to top–down data interpretation and also identifies PTMs within a protein sample (48, 49). Furthermore, deconvolution is crucial for data interpretation, and it is commonly achieved through Xtract, MS-Deconv and YADA (within ProLuCID), among other tools (50). Additionally, in order to give meaning to the identified intact proteins/proteoforms and analyze them more deeply, an integrated network approach can be followed. As an example, "Proteoform" Suite has been recently used for dataset identification and proteoform integration. By assessing its function using gene ontology (GO) analysis, it also enables the visualization of

association and abundance within networks through Cytoscape (51, 52). Although top–down proteomics is still rapidly evolving, the complexity of the analysis and technological issues remain, preventing it to be a typical method to follow when studying high-throughput PTMs.

Middle–down data analysis

MD approaches are based mainly on ESI, where multiple peaks of charged fragment ions are generated. Therefore, it is essential to perform deconvolution prior to MS spectra interpretation. Several tools have been described for this purpose, such as Xtract and YADA (within ProLuCID) or Proteome Discoverer (45, 53). The subsequent dataset analysis and database searches are usually performed using Mascot or Sequest (44, 45, 53). Moreover, new software tools are under development to filter Mascot and Sequest results, such as isoScale, where Mascot results are imported and peptides with confidently assigned combinatorial PTMs are identified, which means that all the modifications are uniquely validated by ions that determine and confirm the localization of a PTM site (45, 53, 54). Since MD proteomics research has a considerable impact on PTM research, specific software tools have been created to analyze PTMs and relevant data. Among these tools are the previously mentioned isoScale software and the Skyline software (55). MD proteomics is still lacking established and standardized tools suitable for data interpretation, and although algorithms and software tools remain under constant development and improvement, this issue is mainly overcome by using TD proteomics tools instead. However, due to a different focus (no proteolytic peptides), such MD analyses are prone to error.

DATA HANDLING AND WORKFLOWS

The general process of data analysis, shown in Figure 3, involves procedures of raw data conversion, deconvolution, normalization, spectral identification, peak alignments, validation, statistical modeling, peptide identification, abundance measurements, protein inference, data storage (raw and processed), data visualization, eventual further data analysis steps such as dataset comparisons and ultimately deposition of data into public data repositories.

Data processing software

A vast amount of computational solutions have been developed to handle and analyze proteomic MS data, ranging in thousands of applications, add-ons and scripts, covering every single aspect of data conversion, deconvolution, normalization and alignment, as listed in website (https://omictools.com/proteomics-category). Currently, the main problem is to find the most appropriate and suitable analysis tool rather than to find a way to analyze the experimental MS spectra. A good overview of the software landscape of such tools can be found in Ref. (56), which also poignantly describes the incompatibility issues when faced with such

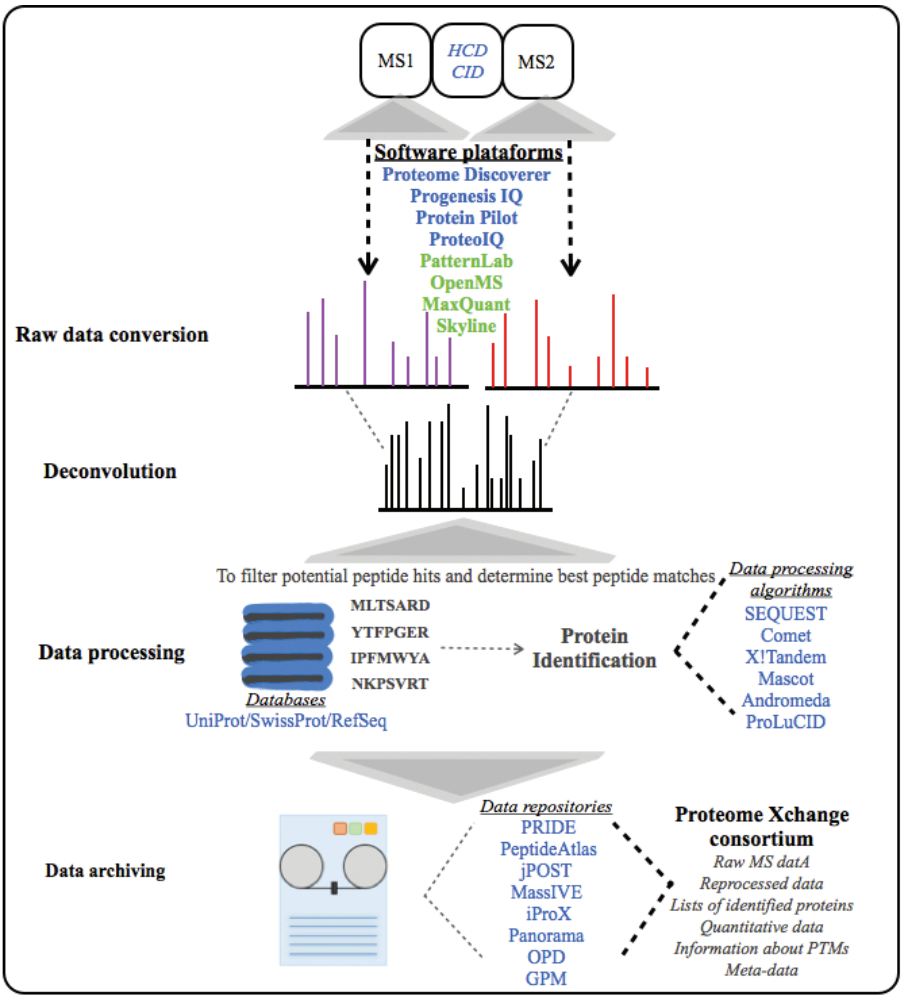

Figure 3 Data processing workflow from raw MS spectra to identified biomolecules. Raw data conversion, deconvolution, data processing and data archiving are the main steps illustrated. The most popular tools within each of them are highlighted in blue. Commercial (blue) and open-source integrated software platforms (green) for the analysis of proteomic data are included. They all encompass modules to manage raw spectral file data, peptide identification using search engines, clustering and sample comparison, identification of PTMs, quantification, statistical analysis and visualization tools. The most common data processing algorithms, data bases and data repositories to release data into the public domain are also shown.

a huge array of computational tools. Therefore, a focused view of the most general, yet commonly used software solutions is summarized in Table 2.

Integrated pipelines

The exuberance of programs and applied algorithms in data processing led to a fragmented landscape of often incompatible steps needed to perform MS data analysis, and the obvious solution was to integrate these various steps into one single workstream implemented in platform tools. A considerable amount of reviews of existing platform software programs are available (57–59). Common amongst many platform solutions is that they usually have one or more of the aforementioned protein/peptide search engines embedded in the workflow. All major MS system manufacturers also provide integrated software solutions for the analysis of proteomic data and specific applications, thereby eliminating the need of having separate software solutions for data acquisition and data processing; however, it needs to be noted that MS instrument control might still require vendor-specific applications. As a consequence, data formats of raw MS data are specific for the manufacturer of the MS equipment, and inter-operability of software solutions is severely hampered and sometimes impossible. This lock-in has understandable commercial reasons, but quite a number of open-source solutions have also been made available over the years. One of the main differences between commercial and open-source solutions is user friendliness, where open-source programs might require specialist computing skills in order to implement the various components of the software programs. However, in recent developments, more user-friendly platforms have been generated that integrate these open-source solutions or algorithms. Therefore, most of these open-source applications feature a modular design, where individual algorithms and procedures are combined to form the entire workflow.

DATA INTERPRETATION AND FUNCTIONAL ANALYSIS

One of the key aspects in proteomic research is the downstream analysis, whereby lists of molecules are interrogated using a variety of software tools in order to put biological meaning into such lists, extract statistically evaluated parameters or match them against other known assemblies (Figure 4). These steps generally involve the use of other databases that hold specific information for each molecule, such as functionality, disease association or pathway data. More than 300 software tools to accomplish various aspects are listed at this website (https://www.ms-utils.org/) alone, and thousands more have been developed and used in proteomics research over the last 20 years. Table 2 lists some of the most common tools used in proteomic downstream analysis.

Statistical approaches

The large-scale nature of proteomic data, which reflects not only the biological factors but also the technical and experimental factors, often requires algorithms to reduce the dimensionality. Statistical tools are an essential part in the analysis of such data, ranging from outlier detection methods and imputation of missing data to expression profiling and group comparisons, including networks and

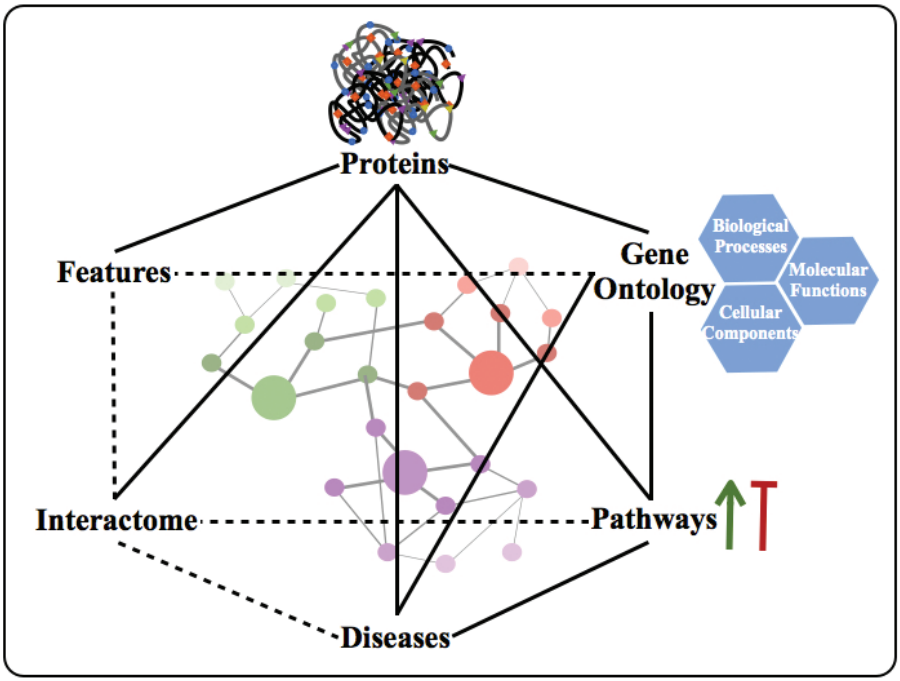

Figure 4 **Downstream data analysis in proteomics research and relationships of analysis scopes.** Full lines depict direct information flow between the analysis or data types, and dashed lines depict indirect associations.

protein cluster detection (60). A vast number of these procedures have been implemented as scripts in the statistical open-source tool R, or in one of its derivatives such as Bioconductor, where a number of packages were written specifically for use in proteomics applications and data analysis (61). A basic first step in the analysis of large MS-derived datasets can also involve a possible enrichment of specific protein families or domains. The InterPro database is an integrated documentation resource for protein domains, families and functional sites, incorporating other databases with similar scope, namely ProDom, PROSITE, PRINTS and Pfam (62). Analysis of the protein landscape using the InterPro resource is generally a practical and efficient way to interrogate proteomic datasets.

Gene ontology

One of the most prominent and heavily used data resource for downstream analysis is the GO database, whose aim is to generate a dynamic, yet controlled vocabulary that can be used in all eukaryotes as the knowledge of gene and protein roles in cells is growing and constantly changing (63). The database unifies similar prior approaches from other databases and describes molecules in terms of their involvement in biological processes, their molecular function and their sub-cellular location in a hierarchical way. Originally, this process of annotating functionality

tags to molecules was done manually, but nowadays it is performed mainly through computational tools. There are many software solutions that make use of the GO database, and surprisingly, depending on the algorithm used in GO-analysis, the results can vary drastically (64). Therefore, one needs to carefully evaluate which tools to use and which are trustworthy in their analysis outcomes.

Pathway analysis

Other high-quality, manually curated databases that extend the knowledge of molecular functionalities are comprised of pathway databases, and more than 600 databases within this scope are currently listed at this website (http://pathguide.org). Pathways, such as signaling and metabolic cascades, can be used to physically link proteins in a concatenated manner to a series of events with a measurable outcome, thereby reducing the complexity of the protein-centric view to a more meaningful one through identification of functional biological processes (65). They can also be used to bridge or integrate data from one omics stream such as proteomics and another like metabolomics. Additionally, many signaling cascades, in particular gene-activation pathways, terminate at the point where gene expression is induced or repressed, thereby breaking the information flow from one signaling event to another via an intermediary step of gene modulation. In order to fill this gap, it is necessary to identify potential transcription factors, their DNA binding sites and the targeted genes (66). Such information can be used for both down-stream pathway mapping and up-stream analysis, thereby enabling the exploration of causes leading to the observed proteomic profile changes, as well as the consequences of such changes.

Interactomes

An additional aspect to consider is that most proteins do not act alone and independently, but rather as an assembly of multiple proteins to perform specific actions by forming transient or stable complexes. Examples are scaffolders that bring proteins into close proximity in protein signaling cascades, protein regulatory networks and structural components. Based on the composition of such complexes, a specific protein might be involved in a function that is fundamentally different from the same molecule participating in an assemblage with other proteins. Therefore, in order to gain a better understanding of the biological data from MS-derived experimentation, the use of protein–protein interaction databases can be particularly helpful (67). Most protein–protein interaction databases contain literature-based interaction data that were manually curated and assessed, whereas some resources use literature mining tools to populate the database, and their data are therefore not necessarily based on experimental observations, but rather predicted interactions.

Disease mapping

Further contextualization of proteomic data can also be achieved by interrogating disease databases, where disease terms are linked to a collection of associated genes derived from the literature. Two such examples are the Online Mendelian

Inheritance in Man (OMIM) (68) and DisGeNET (69). Both are expert-curated databases that analyze text-mined data to establish a link between phenotypes and genes and both have their own web interface to query the databases. While OMIM and its derivative table of gene-disease associations termed MorbidMap are only covering human genes and disease conditions, DisGeNET additionally includes data from animal disease models. Although they are comparable in scope, they both do not use the exact same medical term dictionary, which can cause problems comparing and fusing results using both databases

Integrated frameworks

The diverse nature of biological questions to be answered by proteomics can make it difficult for non-experts in data analysis to make the right selection of analysis tools, and together with specific requirements such as R scripting or programming skills, it can become a daunting endeavor. Yet, new tools are emerging that bring together various data downstream processing procedures most commonly used in omics research such as Babelomics and Cytoscape that will help researchers to put meaning into large-scale datasets, and some of the tools described before have also been integrated into these software solutions as well. Babelomics, although in principle more useful in gene and array analysis, can also be used in the functional characterization of proteomic datasets and other downstream analysis steps (70). It includes a comprehensive suite of modules to perform differential expression profiling, enrichment analysis, GO and pathway analysis, text mining and protein interaction analysis. It is implemented as a web-based application and is freely available and accessible. Cytoscape is an open-source and freely available software framework for interaction network analysis and is offered as a desktop application or a web-plugin (71). In itself, it provides basic functionalities such as graph drawing and network layout and construction and enables linking to large databases. It is extendable by providing a run-time environment for other data analysis plug-ins. Currently, approximately 350 additional apps are available.

CONCLUSION

MS, and in particular the LC-MS/MS shotgun proteomics workflow, is widely used to identify and quantify sample peptides and proteins. The methodology, however, still poses several challenges for large-scale use, such as the MS-manufacturer dependent diverse raw data file formats, the relatively large false-positive peptide assignment rate and the disconnect between observed peptides and originating sample proteins. There are still quite a number of issues to be resolved concerning proteomics in general, such as missing data or data depth, where the sensitivity of the mass spectrometer is insufficient to reliably detect low-abundance molecules, or where the very nature of the molecules under investigation prohibits certain applications, which is commonly encountered with transmembrane spanning proteins. Problems that arise due to masking effects, particularly encountered with high-abundance molecules that raise the detection threshold, are more of a technical issue that can be overcome with improvement

of methodologies, whereas database drift, which is associated with underlying reference databases where accession numbers are lost over time due to various reasons, can pose real problems in the long term.

While many elegant software solutions of data acquisition to spectral data analysis exist, the field is rather fragmented and disjointed when it comes to downstream data analysis such as integrating or merging results derived from pathway mapping, terminology clustering and disease analysis. Yet, tremendous efforts have already begun to pay off in collating and merging individual applications and algorithms into a more cohesive framework. One such framework, the Pan-omics Analysis Database (PADB) initiative, has been in existence for more than 15 years and has been successfully used to address proteomic and genomic large-scale data analysis in various disease areas (72). Another obvious solution is the reuse of existing pipelines and workflows generated in other omics-streams, in particular from the genomics and transcriptomics fields. These tools can be helpful in many ways in proteomics data analysis, yet they might also confuse the picture of available tools and analysis workstreams.

Nevertheless, it is very apparent that since proteomics entered the mainstream and has become an accepted standard in large-scale biological investigations, many breakthroughs were achieved that were unthinkable before. A very new view of the small-scale world has opened and, although the most obvious impact at that moment was how little we understand in terms of molecular flux and interplay, enabled us to start interrogating biological processes on an unprecedented scale. In particular, disease analysis, understanding of abnormal phenotypes and how to pharmacologically interfere with the protein landscape at various stages of disease progression, has started to bear fruit and will continue to do so in the foreseeable future.

Conflict of interest: The authors declare that they have no conflicts of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: We confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s). All original sources have been appropriately acknowledged and/or referenced.

REFERENCES

1. Krishna RG, Wold F. Post-translational modifications of proteins. In: Imahori K, Sakiyama F, editors. Methods in protein sequence analysis. Boston, MA: Springer US; 1993. p. 167–72.
2. Pandey A, Mann M. Proteomics to study genes and genomes. Nature. 2000 Jun 15;405(6788):837–46. http://dx.doi.org/10.1038/35015709
3. Boersema PJ, Kahraman A, Picotti P. Proteomics beyond large-scale protein expression analysis. Curr Opin Biotechnol. 2015 Aug;34:162–70. http://dx.doi.org/10.1016/j.copbio.2015.01.005
4. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: Technologies and their applications. J Chromatogr Sci. 2017 Feb;55(2):182–96. http://dx.doi.org/10.1093/chromsci/bmw167
5. Nilsson T, Mann M, Aebersold R, Yates JR, Bairoch A, Bergeron JJM. Mass spectrometry in high-throughput proteomics: Ready for the big time. Nat Methods. 2010 Sep;7(9):681–5. http://dx.doi.org/10.1038/nmeth0910-681

6. Moore JB, Weeks ME. Proteomics and systems biology: Current and future applications in the nutritional sciences. Adv Nutr. 2011 Jul;2(4):355–64. http://dx.doi.org/10.3945/an.111.000554
7. Carnielli CM, Winck FV, PaesLeme AF. Functional annotation and biological interpretation of proteomics data. Biochim Biophys Acta. 2015 Jan;1854(1):46–54. http://dx.doi.org/10.1016/j.bbapap.2014.10.019
8. Goh WWB, Wong L. Design principles for clinical network-based proteomics. Drug Discov Today. 2016 Jul;21(7):1130–8. http://dx.doi.org/10.1016/j.drudis.2016.05.013
9. Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, et al. Multiplexed MS/MS for improved data-independent acquisition. Nat Methods. 2013 Aug;10(8):744–6. http://dx.doi.org/10.1038/nmeth.2528
10. Xu J, Wang L, Li J. Biological network module-based model for the analysis of differential expression in shotgun proteomics. J Proteome Res. 2014 Dec 5;13(12):5743–50. http://dx.doi.org/10.1021/pr5007203
11. Allmer J. A call for benchmark data in mass spectrometry-based proteomics. J Integr OMICS. 2012;2(2):1–5. http://dx.doi.org/10.5584/jiomics.v2i2.113
12. Colangelo CM, Chung L, Bruce C, Cheung K-H. Review of software tools for design and analysis of large scale MRM proteomic datasets. Methods. 2013 Jun 15;61(3):287–98. http://dx.doi.org/10.1016/j.ymeth.2013.05.004
13. Husi H, Albalat A. Proteomics. In: Padmanabhan S, editor. Handbook of pharmacogenomics and stratified medicine. 1st ed. London: Academic Press; 2014. p. 147–79. http://dx.doi.org/10.1016/B978-0-12-386882-4.00009-8
14. Gutstein HB, Morris JS, Annangudi SP, Sweedler JV. Microproteomics: Analysis of protein diversity in small samples. Mass Spectrom Rev. 2008 Jul–Aug;27(4):316–30. http://dx.doi.org/10.1002/mas.20161
15. Thakur D, Rejtar T, Wang D, Bones J, Cha S, Clodfelder-Miller B, et al. Microproteomic analysis of 10,000 laser captured microdissected breast tumor cells using short-range sodium dodecyl sulfate-polyacrylamide gel electrophoresis and porous layer open tubular liquid chromatography tandem mass spectrometry. J Chromatogr A. 2011 Nov 11;1218(45):8168–74. http://dx.doi.org/10.1016/j.chroma.2011.09.022
16. Feist P, Hummon A. Proteomic challenges: Sample preparation techniques for microgram-quantity protein analysis from biological samples. Int J Mol Sci. 2015 Feb 5;16(2):3537–63. http://dx.doi.org/10.3390/ijms16023537
17. Drissi R, Dubois M-L, Boisvert F-M. Proteomics methods for subcellular proteome analysis. FEBS J. 2013 Nov;280(22):5626–34. http://dx.doi.org/10.1111/febs.12502
18. Shoemaker LD, Kornblum HI. Neural stem cells (NSCs) and proteomics. Mol Cell Proteomics. 2016 Feb;15(2):344–54. http://dx.doi.org/10.1074/mcp.O115.052704
19. Zhao W, Li J, Mills GB. Functional proteomic characterization of cancer cell lines. Oncoscience. 2017 Jun 10;4(5–6):41–2. http://dx.doi.org/10.18632/oncoscience.351
20. Phelan K, May KM. Basic techniques in mammalian cell tissue culture: Basic techniques in mammalian cell tissue culture. Curr Protoc Cell Biol. 2015 Mar 2;66:1.1.1–22. http://dx.doi.org/10.1002/0471143030.cb0101s66
21. Schwaid AG, Krasowka-Zoladek A, Chi A, Cornella-Taracido I. Comparison of the rat and human dorsal root ganglion proteome. Sci Rep. 2018 Sep 7;8(1):13469. http://dx.doi.org/10.1038/s41598-018-31189-9
22. Russo C, Lewis EEL, Flint L, Clench MR. Mass spectrometry imaging of 3D tissue models. Proteomics. 2018 Jul;18(14):e1700462. http://dx.doi.org/10.1002/pmic.201700462
23. Velic A, Macek B, Wagner CA. Toward quantitative proteomics of organ substructures: Implications for renal physiology. Semin Nephrol. 2010 Sep;30(5):487–99. http://dx.doi.org/10.1016/j.semnephrol.2010.07.006
24. Gonneaud A, Asselin C, Boudreau F, Boisvert FM. Phenotypic analysis of organoids by proteomics. Proteomics. 2017 Oct;17(20). http://dx.doi.org/10.1002/pmic.201700023
25. Bendixen E. Animal models for translational proteomics. Proteomics Clin Appl. 2014 Oct;8(9–10):637–9. http://dx.doi.org/10.1002/prca.201470054

26. Skardal A, Murphy SV, Devarasetty M, Mead I, Kang HW, Seol YJ, et al. Multi-tissue interactions in an integrated three-tissue organ-on-a-chip platform. Sci Rep. 2017 Aug 18;7(1):8837. http://dx.doi.org/10.1038/s41598-017-08879-x
27. Khalid N, Arif S, Kobayashi I, Nakajima M. Lab-on-a-chip techniques for high-throughput proteomics and drug discovery. In: Santos HA, Dongfei Liu D, Zhang H, editors. Micro and nano technologies, microfluidics for pharmaceutical applications. Norwich, NY: William Andrew Publishing; 2019. p. 371–422. https://doi.org/10.1016/B978-0-12-812659-2.00014-4
28. Raimondo F, Morosi L, Chinello C, Magni F, Pitto M. Advances in membranous vesicle and exosome proteomics improving biological understanding and biomarker discovery. Proteomics. 2011 Feb;11(4):709–20. http://dx.doi.org/10.1002/pmic.201000422
29. Guay C, Regazzi R. Exosomes as new players in metabolic organ cross-talk. Diabetes Obes Metab. 2017 Sep;19(Suppl 1):137–46. http://dx.doi.org/10.1111/dom.13027
30. Kwasnik A, Tonry C, Ardle AM, Butt AQ, Inzitari R, Pennington SR. Proteomes, their compositions and their sources. In: Mirzaei H, Carrasco M, editors. Modern proteomics – Sample preparation, analysis and practical applications. Cham: Springer International Publishing; 2016. p. 3–21.
31. Bladergroen MR, van der Burgt YEM. Solid-phase extraction strategies to surmount body fluid sample complexity in high-throughput mass spectrometry-based proteomics. J Anal Methods Chem. 2015;2015:250131. http://dx.doi.org/10.1155/2015/250131
32. Hsieh S-Y, Chen R-K, Pan Y-H, Lee H-L. Systematical evaluation of the effects of sample collection procedures on low-molecular-weight serum/plasma proteome profiling. Proteomics. 2006 May;6(10):3189–98. http://dx.doi.org/10.1002/pmic.200500535
33. Sköld K, Alm H, Scholz B. The impact of biosampling procedures on molecular data interpretation. Mol Cell Proteomics. 2013 Jun;12(6):1489–501. http://dx.doi.org/10.1074/mcp.R112.024869
34. Kelstrup CD, Bekker-Jensen DB, Arrey TN, Hogrebe A, Harder A, Olsen JV. Performance evaluation of the Q exactive HF-X for shotgun proteomics. J Proteome Res. 2018 Jan 5;17(1):727–38. http://dx.doi.org/10.1021/acs.jproteome.7b00602
35. Sadygov RG, Cociorva D, Yates JR. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. Nat Methods. 2004 Dec;1(3):195–202. http://dx.doi.org/10.1038/nmeth725
36. Pandeswari PB, Sabareesh V. Middle-down approach: A choice to sequence and characterize proteins/proteomes by mass spectrometry. RSC Adv. 2019 Jan 2;9:313–44. http://dx.doi.org/10.1039/C8RA07200K
37. Zhang Q, Ma C, Gearing M, Wang PG, Chin L-S, Li L. Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer's disease. Acta Neuropathol Commun. 2018 Mar 1;6(1):19. http://dx.doi.org/10.1186/s40478-018-0524-2
38. Yu Y, Bekele S, Pieper R. Quick 96FASP for high throughput quantitative proteome analysis. J Proteomics. 2017 Aug 23;166:1–7. http://dx.doi.org/10.1016/j.jprot.2017.06.019
39. Spanka D-T, Konzer A, Edelmann D, Berghoff BA. High-throughput proteomics identifies proteins with importance to postantibiotic recovery in depolarized persister cells. Front Microbiol. 2019 Mar 6;10:378. http://dx.doi.org/10.3389/fmicb.2019.00378
40. Dams M, Dores-Sousa JL, Lamers R-J, Treumann A, Eeltink S. High-resolution nano-liquid chromatography with tandem mass spectrometric detection for the bottom-up analysis of complex proteomic samples. Chromatographia. 2018 Nov 7;82(1):101–10. http://dx.doi.org/10.1007/s10337-018-3647-5
41. Sinitcyn P, Rudolph JD, Cox J. Computational methods for understanding mass spectrometry–based shotgun proteomics data. Annu Rev Biomed Data Sci. 2018 Jul;1:207–34. http://dx.doi.org/10.1146/annurev-biodatasci-080917-013516
42. Thomas S, Hao L, Ricke WA, Li L. Biomarker discovery in mass spectrometry-based urinary proteomics. Proteomics Clin Appl. 2016 Apr;10(4):358–70. http://dx.doi.org/10.1002/prca.201500102
43. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017 Jan 4;45(D1):D362–8. http://dx.doi.org/10.1093/nar/gkw937
44. Cristobal A, Marino F, Post H, van den Toorn HW, Mohammed S, Heck AJ. Toward an optimized workflow for middle-down proteomics. Anal Chem. 2017 Mar 21;89(6):3318–3325. http://dx.doi.org/10.1021/acs.analchem.6b03756

45. Greer SM, Sidoli S, Coradin M, Schack Jespersen M, Schwämmle V, Jensen ON, et al. Extensive characterization of heavily modified histone tails by 193 nm ultraviolet photodissociation mass spectrometry via a middle-down strategy. Anal Chem. 2018 Sep 4;90(17):10425–33. http://dx.doi.org/10.1021/acs.analchem.8b02320
46. Zhang H, Ge Y. Comprehensive analysis of protein modifications by top-down mass spectrometry. Circ Cardiovasc Genet. 2011 Dec;4(6):711. http://dx.doi.org/10.1161/CIRCGENETICS.110.957829
47. McCool EN, Chen D, Li W, Liu Y, Sun LL. Capillary zone electrophoresis-tandem mass spectrometry with ultraviolet photodissociation (213 nm) for large-scale top–down proteomics. Anal Methods. 2019 May 7;11:2855–61. http://dx.doi.org/10.1039/C9AY00585D
48. Cleland TP, DeHart CJ, Fellers RT, VanNispen AJ, Greer JB, LeDuc RD, et al. High-throughput analysis of intact human proteins using UVPD and HCD on an orbitrap mass spectrometer. J Proteome Res. 2017 May 5;16(5):2072–9. http://dx.doi.org/10.1021/acs.jproteome.7b00043
49. Skinner OS, Haverland NA, Fornelli L, Melani RD, Do Vale LHF, Seckler HS, et al. Top-down characterization of endogenous protein complexes with native proteomics. Nat Chem Biol. 2018 Jan;14(1):36–41. http://dx.doi.org/10.1038/nchembio.2515
50. Tholey A, Becker A. Top-down proteomics for the analysis of proteolytic events – Methods, applications and perspectives. Biochim Biophys Acta Mol Cell Res. 2017 Nov;1864(11 Pt B):2191–9. http://dx.doi.org/10.1016/j.bbamcr.2017.07.002
51. Cesnik AJ, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, et al. Proteoform suite: Software for constructing, quantifying, and visualizing proteoform families. J Proteome Res. 2018 Jan 5;17(1):568–78. http://dx.doi.org/10.1021/acs.jproteome.7b00685
52. Schaffer LV, Rensvold JW, Shortreed MR, Cesnik AJ, Jochem A, Scalf M, et al. Identification and quantification of murine mitochondrial proteoforms using an integrated top-down and intact-mass strategy. J Proteome Res. 2018 Oct 5;17(10):3526–36. http://dx.doi.org/10.1021/acs.jproteome.8b00469
53. Sidoli S, Lu C, Coradin M, Wang X, Karch KR, Ruminowicz C, et al. Metabolic labeling in middle-down proteomics allows for investigation of the dynamics of the histone code. Epigenetics Chromatin. 2017 Jul 6;10(1):34. http://dx.doi.org/10.1186/s13072-017-0139-z
54. Sidoli S, Garcia BA. Middle-down proteomics: A still unexploited resource for chromatin biology. Expert Rev Proteomics. 2017 Jul;14(7):617–26. http://dx.doi.org/10.1080/14789450.2017.1345632
55. Moradian, A, Franco C, Sweredoski, MJ, Hess, S. Middle-down electron capture dissociation and electron transfer dissociation for histone analysis. J Anal Sci Technol. 2015 Dec;6:21. http://dx.doi.org/10.1186/s40543-015-0060-7
56. Krappmann M, Luthardt M, Lesske F, Letzel T. The software-landscape in (prote)omic research. J Proteomics Bioinform. 2015;8(7):164–75. http://dx.doi.org/10.4172/jpb.1000365
57. Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Röst HL, et al. A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol. 2016 Nov;34(11):1130–6. http://dx.doi.org/10.1038/nbt.3685
58. Välikangas T, Suomi T, Elo LL. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief Bioinform. 2018 Nov 27;19(6):1344–55.
59. Chen Y, Wang F, Xu F, Yang T. Mass spectrometry-based protein quantification. Adv Exp Med Biol. 2016;919:255–79. http://dx.doi.org/10.1007/978-3-319-41448-5_15
60. Urfer W, Grzegorczyk M, Jung K. Statistics for proteomics: A review of tools for analyzing experimental data. Proteomics. 2006 Sep;6(Suppl 2):48–55. http://dx.doi.org/10.1002/pmic.200600554
61. Gatto L, Christoforou A. Using R and bioconductor for proteomics data analysis. Biochim Biophys Acta. 2014 Jan;1844(1 Pt A):42–51. http://dx.doi.org/10.1016/j.bbapap.2013.04.032
62. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001 Jan 1;29(1):37–40. http://dx.doi.org/10.1093/nar/29.1.37
63. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9. http://dx.doi.org/10.1038/75556
64. Khatri P, Drăghici S. Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics. 2005 Sep 15;21(18):3587–95. http://dx.doi.org/10.1093/bioinformatics/bti565

65. Wu X, Hasan MA, Chen JY. Pathway and network analysis in proteomics. J Theor Biol. 2014 Dec 7;362:44–52. http://dx.doi.org/10.1016/j.jtbi.2014.05.031
66. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D91–4. http://dx.doi.org/10.1093/nar/gkh012
67. Koh GC, Porras P, Aranda B, Hermjakob H, Orchard SE. Analyzing protein-protein interaction networks. J Proteome Res. 2012 Apr 6;11(4):2014–31. http://dx.doi.org/10.1021/pr201211w
68. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan;43(Database issue):D789–98. http://dx.doi.org/10.1093/nar/gku1205
69. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015 Apr 15;2015:bav028. http://dx.doi.org/10.1093/database/bav028
70. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, et al. Babelomics: An integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W210–13. http://dx.doi.org/10.1093/nar/gkq388
71. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504. http://dx.doi.org/10.1101/gr.1239303
72. Husi H. NMDA receptors, neural pathways, and protein interaction databases. Int Rev Neurobiol. 2004;61:49–77. http://dx.doi.org/10.1016/S0074-7742(04)61003-8