A Practical Beginner's Guide to Proteomics

This manuscript (permalink) was automatically generated from jessegmeyerlab/proteomics-tutorial@9622b48 on December 19, 2022.

Authors

- **Benjamin A. Neely**
  0000-0001-6120-7695 · neely · neely615
  Chemical Sciences Division, National Institute of Standards and Technology, NIST Charleston · Funded by NIST

- **Amit Kumar Yadav**
  0000-0002-9445-8156 · aky · theoneamit
  Translational Health Science and Technology Institute · Funded by Grant BT/PR16456/BID/7/624/2016 (Department of Biotechnology, India); Grant Translational Research Program (TRP) at THSTI funded by DBT

- **Emma H. Doud**
  0000-0003-0049-0073 · edoud1 · fireinlab
  Center for Proteome Analysis, Indiana University School of Medicine, Indianapolis, Indiana, USA

- **Dina Schuster**
  0000-0001-6611-8237 · dschust-r · dina_sch
  Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland; Department of Biology, Institute of Molecular Biology and Biophysics, ETH Zurich, Zurich 8093, Switzerland; Laboratory of Biomolecular Research, Division of Biology and Chemistry, Paul Scherrer Institute, Villigen 5232, Switzerland

- **Martín L. Mayta**
  0000-0002-7986-4551 · martinmayta · MartinMayta2
  School of Medicine and Health Sciences, Center for Health Sciences Research, Universidad Adventista del Plata, Libertador San Martín 3103, Argentina; Molecular Biology Department, School of Pharmacy and Biochemistry, Universidad Nacional de Rosario, Rosario 2000, Argentina

- **Jesse G. Meyer**
  0000-0003-2753-3926 · jessegmeyerlab · j_my_sci
  Department of Computational Biomedicine, Cedars Sinai Medical Center · Funded by Grant R21 AG074234; Grant R35 GM142502

- **Muralidharan Vanuopadath**
  0000-0002-9364-917X · vanuopadathmurali · V_MuraleeDhar
  School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam-690 525, Kerala, India

- **Devasahayam Arokia Balaya Rex**
  0000-0002-9556-3150 · ArokiaRex · rexprem
  Center for Systems Biology and Molecular Medicine, Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore 575018, India

Abstract

Proteomics is the large scale study of protein structure and function from biological systems. “Shotgun proteomics” or “bottom-up proteomics” is the prevailing strategy, in which proteins are hydrolyzed into peptide that are analyzed by mass spectrometry. Proteomics studies can be applied to diverse studies ranging from simple protein identification to studies of protein-protein interactions, absolute and relative protein quantification, post-translational modifications, and protein stability. To enable this range of different experiments, there are diverse strategies for proteome analysis. The nuances of how proteomic workflows differ may be difficult to understand for new practitioners. Here, we provide a comprehensive tutorial of different proteomics methods. Our tutorial covers all necessary steps starting from protein extraction and ending with biological interpretation. We expect that this work will serve as a basic resource for new practitioners of the field of shotgun or bottom-up proteomics.

Introduction

Proteomics is the large scale study of protein structure and function. Proteins are translated from mRNAs that are transcribed from the genome. Although the genome encodes potential cellular functions and states, the study of proteins is necessary to truly understand biology. Currently, proteomic studies are facilitated by mass spectrometry, although alternative methods are being developed.

Modern proteomics started around the year 1990 with the introduction of soft ionization methods that enabled, for the first time, transfer of large biomolecules into the gas phase without destroying them [1,2]. Shortly afterward, the first computer algorithm for matching peptides to a database was introduced [**PMID:24226387?**]. Another major milestone that allowed identification of over 1000 proteins were actually improvements to chromatography [3]. As the volume of data exploded, methods for statistical analysis transitioned use from the wild west to modern informatics based on statistical models [4] and the false discovery rate [5].

Two strategies of mass spectrometry-based proteomics differ fundamentally by whether proteins are cleaved into peptides before analysis: “top-down” and “bottom-up”. Bottom-up proteomics (also refered to as shotgun proteomics) is defined by the hydrolysis of proteins into peptide pieces [6]. Therefore, bottom-up proteomics does not actually measure proteins, but must infer their presence [4]. Sometimes proteins are infered from only one peptide sequence representing a small fraction of the total protein sequence predicted from the genome. In contrast, top-down proteomics attempts to measure all proteins intact [7]. The potential benefit of top-down proteomics is the ability to measure proteoforms [8]. However, due to analytical challenges, the depth of protein coverage that is achievable by top-down proteomics is less than the depth that is achievable by bottom-up proteomics.

In this tutorial we focus on the bottom-up proteomics workflow. The most common version of this workflow is generally comprised of the following steps. First, proteins in a biological sample must be extracted. Usually this is done by denaturing and solubilizing the proteins while disrupting DNA and tissue. Next, proteins are hydrolyzed into peptides, usually using a protease like trypsin. Peptides from proteome hydrolysis must be purified. Most often this is done with reversed phase chromatography cartridges or tips. The peptides are then almost always separated by liquid chromatography before they are ionized and introduced into a mass spectrometer. The mass spectrometer then collects precursor and fragment ion data from those peptides. The data analysis is usually the rate limiting step. Peptides must be identified, and proteins are inferred and quantities are assigned. Changes in proteins across conditions are determined with statistical tests, and results must be interpreted in the context of the relevant biology.

There are many variations on this workflow. The wide variety of experimental goals that are achievable with proteomics technology leads to a wide variety of potential proteomics workflows. Even choice is important and every choice will affect the results. In this tutorial, we cover all of the required steps in detail to serve as a tutorial for new proteomics practioners.

1. Biochemistry basics
2. Types of experiments enabled by proteomics
3. Protein extraction
4. proteolysis
5. Isotopic Labeling
6. Enrichments
7. Peptide purification
8. Mass Spectrometry

9. Peptide Ionization
10. Data Acquisition
11. Basic Data Analysis
12. Biological Interpretation
13. Sample Fractionation
14. Tandem MS
15. Experimental considerations and design

Biochemistry Basics

Proteins

Proteins are large biomolecules or biopolymers made up of amino acids which are linked by peptide bonds. They perform various functions in living organisms ranging from having structural roles to functional involvement in cellular signaling and the catalysis of chemical reactions (enzymes). Proteins are made up of 20 different amino acids (not counting pyrrolysine and selenocysteine, which only occur in specific organisms) and their sequence is encoded in their corresponding genes. The human genome encodes more than 20,000 different proteins. Each protein is present at a different abundances. Previous studies have shown that the concentration range of proteins can span over a range of at least seven orders of magnitude to up to 20 000 000 copies per cell and that their distribution is tissue-specific[[9]][10]. Due to genetic variations, as well as alternative splicing and post-translational modifications, multiple different proteoforms can be produced from one single gene[11].

PTMs

Proteins can be post-translationally modified through enzymatic and non-enzymatic reactions *in vivo* and *in vitro* [12]. Post-translational modifications (PTMs) can be reversible or non-reversible covalent modifications of amino acids. The most commonly studied and biologically relevant post-translational modifications include ubiquitinations (Lys, Cys, Ser, Thr, N-term), succinylations (Lys), methylations (Arg, Lys, His, Glu, Asn, Cys), disulfide bonds (Cys-Cys), oxidations (Met, Trp, His, Cys), phosphorylations (Ser, Thr, Tyr, His), acetylations (Lys, N-term), glycosylations (Arg, Asp, Cys, Ser, Thr, Tyr, Trp) and lipidations. Post-translational modification of a protein can alter its function, activity, structure, location and interactions. Deregulation of PTMs is linked to cellular stress and diseases[13].

Protein Structure

Almost all proteins (except for intrinsically disordered proteins[14]) fold into 3D structures either by themselves or assisted through chaperones[15]. There are four levels relevant to the folding of any protein:

- Primary structure: The protein's linear amino acid sequence, with amino acids connected through peptide bonds.
- Secondary structure: The amino acid chain's folding: α-helix, β-sheet or turn.
- Tertiary structure: The three-dimensional structure of the protein.
- Quarternary structure: The structure of several protein molecules/subunits in one complex.

Biochemical and biophysical techniques for studying complex protein mixtures

Concentration/total protein amount determination

- UV-Vis Spectrophotometry (Bradford assay, BCA assay, Lowry, Folin)

Electrophoresis

- SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis)

- 2D-PAGE
- Isoelectric focussing (IEF)
- Western blotting

Biochemical and biophysical analytical techniques for studying purified proteins

Concentration determination

- UV-Vis Spectrophotometry

Protein size and aggregation

- Dynamic light scattering (DLS)
- Size-exclusion chromatography (SEC) or size-exclusion with multi angle laser light scattering (SEC-MALLS)
- SDS-PAGE
- Mass photometry

Structural techniques

- Circular dichroism (CD) spectropolarimetry
- X-ray crystallography
- Protein nuclear magnetic resonance spectroscopy (NMR)
- (cryo-)Electron microscopy (EM)

Interactions and binding kinetics

- Surface plasmon resonance (SPR)
- Isothermal titration calorimetry (ITC)

Buffers

Types of Experiments

A wide range of questions are addressable with proteomics technology, which translates to a wide range of variations of proteomics workflows. Sometimes identifying what proteins are present is desired, and sometimes the quantities of as many proteins as possible are desired. Proteomics experiments can be both qualitative and quantitative.

Qualitative experiments

- Identifying proteins
- Identifying post translational modifications
- Identifying protein isoforms

Quantitative experiments

- Protein abundance changes
- Phosphoproteomics
- Glycoproteomics
- Structural techniques (XL-MS, HDX-MS, FPOP, protein-painting, LiP-MS, radical footprinting, ion mobility)
- Protein stability and small molecule binding (Thermal proteome profiling, TPP, or cellular thermal shift assay, CETSA)
- Protein-protein interactions (PPIs): AP-MS, APEX, BioID

Protein Extraction

Protein extraction from the sample of interest is the initial phase of any mass spectrometry-based proteomics experiment. Thought should be given to any planned downstream assays, specific needs of proteolysis (LiP-MS, post translational modification enrichments, enzymatic reactions, glycan purification or hydrogen-deuterium exchange experiments) long term project goals (reproducibility, multiple sample types, low abundance samples) as well as to the initial experimental question (coverage of a specific protein, subcellular proteomics, global proteomics, protein-protein interactions or immune or affinity enrichment of a specific classes of modifications.) The 2009 version of Methods in Enzymology: guide to Protein Purification [16] serves as a deep dive into how molecular biologists and biochemists traditionally thought about protein extraction. The Protein Protocols handbook [17] and the excellent review by Linn [**PMID:19892162?**] are good sources of general proteomics protocols for a scientist new to the field. Any change in extraction conditions should be expected to create potential changes in downstream results. Be sure to think about and optimize the protein extraction step first and stick with a protocol that works for your needs. If a collaborator is attempting to reproduce your results, make sure they begin with the same extraction protocols.

Buffer choice

General proteomics

A common question to proteomics core facilities is, "What is the best buffer for protein extraction?" Unfortunately, there is no one correct answer. For global proteomics experiments where maximizing the number of protein or peptide identifications is a goal, a buffer of neutral pH (50-100 mM PBS, Tris, HEPES, ammonium bicarbonate, triethanolamine bicarbonate; pH 7.5-8.5) is used in conjunction with a chaotrope or surfactant to denature and solubilize proteins (e.g., 8 M urea, 6 M guanidine, 5% SDS) [**PMID:16152629?**,**PMID:20722421?**]. Often other salts like 50-150 mM NaCl are also added. Complete denaturation of the proteins in the sample in a timely fashion is an advantage as it generally prevents changes to protein status by endogenous proteases, kinases, phosphotases, and other enzymes. If intact protein separations are planned (based on size or isoelectric point) choose a denaturant compatible with those methods, such as SDS[**PMID:31249407?**]. Compatibility with protease (typically trypsin) and peptide cleanup steps will need to be considered. 8 M urea must be diluted to 2 M or less for trypsin and chymotrypsin digestions, while guanidine and SDS should be removed either through protein precipitation, through filter-assisted sample preparation (FASP), or similar solid phase digestion techniques. Note that some buffers can potentially introduce modifications onto proteins such as carbamylation from urea at high temperatures [**PMID:24161613?**].

Protein-protein interactions

Denaturing conditions will efficiently extract proteins – but they will denature/disrupt most protein-protein interactions. If you are working on an immune- or affinity purification of a specific protein and expect to analyze enzymatic activity, structural features, and/or protein-protein interactions, a non-denaturing lysis buffer should be utilized [**PMID:21364760?**,**PMID:10504710?**]. Check the calculated pI and hydrophobicity (the Expasy.org resource ProtParam is useful fo this) for a good idea of starting pH/conductivity, but you may need to perform a stability screen. In general, a good starting point for the buffer will still be close to neutral pH with 50-250 mM NaCl, but specific proteins may require pH as low as 2 or as high as 9 for stable extraction. A low percent of mass spec compatible detergent may also be used. Newer mass spectrometry compatible detergents are also useful for protein extraction and ease of downstream processing – including Rapigest® (Waters), N-octyl-β-glucopyranoside, Azo [**PMID:33232116?**], PPS silent surfactant [**PMID:21280217?**], sodium laurate [**PMID:23555778?**], and

sodium deoxycholate[PMID:17022626?]. AVOID the use of tween-20, triton-X, NP-40, and PEGs as these compounds are challenging to remove after digestion [PMID:29726681?].

Optional additives

For non-denaturing buffer conditions, which preserve tertiary and quaternary protein structures, additional additives may not be necessary for successful extraction and to prevent proteolysis or PTM modifications throughout the extraction process. Protease, phosphatase and deubiquitinase inhibitors are optional additives in less denaturing conditions or in experiments focused on specific post-translational modifications. Keep in mind that protease inhibitors may impact digestion conditions and will need to be diluted or removed prior to trypsin addition. For extraction of DNA or RNA binding proteins, addition of a small amount of nuclease or benzonase might be useful for degradation of any bound nucleic acids and result in a more consistent digestion [PMID:23792921?].

Mechanical or Sonic Disruption

Cell lysis

One typical lysis buffer is 8 M urea in 100 mM Tris, pH 8.5; the pH based on optimum trypsin activity [PMID:25664860?] Small mammalian cell pellets and exosomes will lyse almost instantly upon addition denaturing buffer. If non-denaturing conditions are desired, osmotic swelling and subsequent shearing or sonication can be applied [18]. Efficiency of extraction and degradation of nucleic acids can be improved using various sonication methods: 1) probe sonicator with ice; 2) water bath sonicator with ice or cooling; 3) bioruptor® sonication device 4) Adaptive focused acoustics (AFA®) [PMID:21060726?]. Key to these additional lysis techniques are to keep the temperature of the sample from rising significantly which can cause proteins to aggregate or degrade. Some cell types may require additional force for effective lysis (see below). For cells with cell walls (i.e. bacteria or yeast), lysozyme is often added in the lysis buffer. Any added protein will be present in downstream results, however, so excessive addition of lysozyme is to be avoided unless tagged protein purification will occur.

Tissue/other lysis

Although small pieces of soft tissue can often be successfully extracted with the probe and sonication methods described above, larger/harder tissues as well as plants/yeast/fungi are better extracted with some form of additional mechanical force. If proteins are to be extracted from a large amount of sample, such as soil, feces, or other diffuse input, one option is to use a dedicated blender and filter the sample, followed by centrifugation. If samples are smaller, such as tissue, tumors, etc., cryo-homogenization is recommended. The simplest form of this is grinding the sample with liquid nitrogen and a mortar and pestle. Tools such as bead beaters (i.e. FastPrep-24®) are also used, where the sample is placed in a tube with appropriately sized glass or ceramics beads and shaken rapidly. Cryo-mills are chambers where liquid nitrogen is applied around a vessel and large bead or beads. Cryo-fractionators homogenize samples in special bags that are frozen in liquid nitrogen and smashed with various degrees of force [PMID:34002278?]. After homogenization, samples can be sonicated by one of the methods above to fragment DNA and increase solubilization of proteins.

Measuring the efficiency of protein extraction

Following protein extraction, samples should be centrifuged (10-14,000 g for 10-30 min depending on sample type) to remove debris and any unlysed material prior to determinining protein concentration. The amount of remaining insoluble material should be noted throughout an experiment as a large change may indicate protein extraction issues. Protein concentration can be calculated using a

number of assays or tools [PMID:18429326?,PMID:12703310?]; generally absorbance measuremnts are facile, fast and affordable, such as Bradford or BCA assays. Protein can also be estimated by tryptophan fluorescence, which has the benefit of not consuming sample [19]. A nanodrop UV spectrophotometer may be used to measure absorbance at UV280. Consistency in this method is important as each method will have inherent bias and error [PMID:26342307?,PMID:30234128?]. Extraction buffer components will need to be compatible with any assay chosen; alternatively, buffer may be removed (see below) prior to protein concentration calculation.

Reduction and alkylation

Typically, disulfide bonds in proteins are reduced and alkylated prior to proteolysis in order to disrupt structures and simplify peptide analysis. This allows better access to all residues during proteolysis and removes the crosslinked peptides created by S-S inter peptide linkages. There are a variety of reagent options for these steps. For reduction, the typical agents used are 5-15 mM concentration of tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl), dithiothreitol (DTT), or 2-mercaptoethanol (2BME). TCEP-HCl is an efficient reducing agent, but it also significantly lowers sample pH, which can be abated by increasing sample buffer concentration or resuspending TCEP-HCl in an appropriate buffer system (i.e 1M HEPES pH 7.5).
Following the reducing step, a slightly higher 10-20mM concentration of alkylating agent such as chloroacetamide/iodoacetamide or n-ethyl maleimide is used to cap the free thiols [PMID:29019370; [PMID:15351294?]; [PMID:28539326?]]. In order to monitor which cysteine residues are linked or modified in a protein, it is also possible to alkylate free cysteines with one reagent, reduce di-sulfide bonds (or other cysteine modifications) and alkylate with a different reagent [PMID:32132231?,PMID:28445428?,PMID:23074338?]. Alkylation reactions are generally carried out in the dark at room temperature to avoid excessive off-target alkylation of other amino acids.

Removal of buffer/interfering small molecules

If extraction must take place in a buffer which is incompatible for efficient proteolysis (check the guidelines for the protease of choice), then protein cleanup should occur prior to digestion. This is generally performed through precipitation of proteins. The most common types are 1) acetone, 2) trichloroacetic acid (TCA), and 3) methanol/chloroform/water [PMID:14753699?,PMID:19892180?]. Proteins are generally insoluble in most pure organic solvents, so cold ethanol or methanol are sometimes used. Pellets should be washed with organic solvent for complete removal especially of detergents. Alternatively, solid phase based digestion methods such as S-trap [PMID:33750040?], FASP [PMID:19377485?,PMID:30259475?], SP3 [PMID:3117935?,PMID:28948796?] and on column/bead can allow for proteins to be applied to a solid phase and buffers removed prior to proteolysis [PMID? 29754492]. Specialty detergent removal columns exist (Pierce/Thermo Fisher Scientific) but add expense and time consuming steps to the process. Relatively low concentrations of specific detergents, such as 1% deoxycholate (DOC), or chaotropes (i.e. 1M urea) are compatible with proteolysis by trypsin/Lys-C. Often proteolysis-compatible concentrations of these detergents and chaotropes are achieved by diluting the sample in appropriate buffer (i.e. 100 mM ammonium bicarbonate, pH 8.5) after cell or tissue lysis in a higher concentration. DOC can then be easily removed by precipitation or phase separation [PMID:18183947?] following digestion by acidification of the sample to pH 2-3.
Any small-molecule removal protocol should be tested for efficiency prior to implementing in a workflow with many samples as avoiding detergent (or polymer) contamination in the LC/MS is very important.

Protein quantification

After proteins are isolated from the sample matrix, they are often quantified. Protein quantification is important to assess the yeild of an extraction procedure, and to adjust the scale of the downstream processing steps to match the amount of protein. For example, when purifying peptides, the amount of sorbent should match the amount of material to be bound. Presently, there is a wide variety of techniques to quantitate the amount of protein present in a given sample. These methods can be broadly divided into three types as follows:

Colorimetry-based methods:

The method includes different assays like Coomassie Blue G-250 dye binding (the Bradford assay), the Folin-Lowry assay, the bicinchoninic acid (BCA) assay and the biuret assay [PMID:10075906?]. The most commonly used method is the BCA assay. In the BCA method the peptide bonds of the protein reduce cupric ions [Cu2+] to cuprous ions [Cu+] at a rate which is proportional to the amount of protein present in the sample. Subsequently, the BCA reagent binds to the cuprous ions, leading to the formation of a complex which absorbs 562 nm wavelength light. This permits a direct correlation between sample protein concentration and absorbance [PMID:3843705?,PMID:7951748?]. The Bradford assay is another method for protein quantification also based on colorimetry principle. It relies on the interaction between the Coomassie brilliant blue dye and the protein based on hydrophobic and electrostatic interactions. Dye binding shifts the absorption maxima from 470 nm to 595 nm [PMID:32238597?]. Similarly, the Folin- Lowry method is a two-step colorimetric assay. Step one is the biuret reaction wherein complexes of copper with the nitrogen in the protein molecule are formed. In the second step, the complexed tyrosine and tryptophan amino acids react with Folin–Ciocalteu phenol reagent generating an intense, blue-green color absorbing light at 650–750 nm [PMID:6744121?].

Another simple but less reliable protein quantification method of UV-Vis Absorbance at 280 nm estimates the protein concentration by measuring the absorption of the aromatic residues; tyrosine, and tryptophan, at 280 nm [PMID:34533299?].

Fluorescence-based methods:

Colorimetric assays are inexpensive and require common lab equipment, but colorimetric detection is less sensitive than fluorescence. Protein in proteomic samples can be quantified using intrinsic fluorescence of tryptophan based on the assumption that approximately 1% of all amino acids in the proteome are tryptophan [PMID:25837572?].

NanoOrange is an assay for the quantitative measurement of proteins in solution using the NanoOrange reagent, a merocyanine dye that produces a large increase in fluorescence quantum yield when it interacts with detergent-coated proteins. Fluorescence is measured using 485-nm excitation and 590-nm emission wavelengths. The NanoOrange assay can be performed using fluorescence microplate readers, fluorometers, and laser scanners that are standard in the laboratory [PMID:12703310?].

3-(4-carboxybenzoyl)quinoline-2-carboxaldehyde (CBQCA) is a sensitive fluorogenic reagent for amine detection, which can be used for analyzing proteins in solution. As the number of accessible amines in a protein is modulated by its concentration, CBQCA has a greater sensitivity and dynamic range when measuring protein concentration [PMID:9025944?].

Summary

Often you will be given protein extraction conditions from molecular biologists or biochemistry which you will have to make work with downstream mass spectrometry applications. For bottom-up

proteomics, the overreaching goal is efficient and consistent extraction and digestion. A range of mechanical and non-mechanical extraction protocols have been developed and the use any one specific technique is generally dictated by sample type or assay requirements (i.e. native versus non-native extraction). Extraction can be aided by the addition of detergents and/or chaotropes to the sample, but care should be taken that these additives do not interfere with the sample digestion step or downstream mass-spectrometry applications.

Proteolysis

Proteolysis is the defining step that differentiates bottom-up or shotgun proteomics from top-down proteomics. Hydrolysis of proteins is extremely important because it defines the population of potentially identifyable peptides. Generally peptides between a length of 7-35 amino acids are considered useful for mass spectrometry analysis. Peptides that are too long are difficult to identify by tandem mass spectrometry, or may be lost during sample preparation due to irreversible binding with solid-phase extraction sorbents. Peptides that are too short are also not useful because they may match to many proteins during protein inference. There are many choices of enzymes and chemicals that hydrolyze proteins into peptides. This section summarizes potential choices and their strengths and weaknesses.

Trypsin is the most common choice of protease for proteome hydrolysis [20]. Trypsin is favorable because of its specificity, availability, efficiency and low cost. Trypsin cleaves at the C-terminus of basic amino acids, Arg and Lys. Many of the peptides generated from trypsin are short in length (less than ~ 20 amino acids), which is ideal for chromatographic separation, MS-based peptide fragmentation and identification by database search. The main drawback of trypsin is that majority (56%) of the tryptic peptides are ≤ 6 amino acids, and hence using trypsin alone limits the observable proteome [**PMID:20113005?**,**PMID:25823410?**,**PMID:30687733?**]. This limits the number of identifiable protein isoforms and post-translational modifications.

3. theoretical studies of proteolysis enzymes [21]
4. Challenges associated with alternative enzyme choices (non-specific and semi-specific enzymes)

Many alternative proteases are available with different specificities that complement trypsin to reveal different proteomic sequences [**PMID:12643544?**,**PMID:20113005?**], which can help distinguish protein isoforms [**PMID:27123950?**]. The enzyme choice mostly depends on the application. In general, for a mere protein identification mostly trypsin is the choice due to the reasons aforementioned. However, alternative enzymes can facilitate *de novo* assembly when the genomic data information is limited in the public database repositories [22,23,**pmid:31615963?**,**pmid:30622160?**,**pmid:29990557?**]. Use of multiple proteases for proteome digestion also can improve the sensitivity and accuracy of protein quantification [**PMID:30336047?**]. Moreover, by providing an increased peptide diversity, the use of multiple proteases can expand sequence coverage and increase the probability of finding peptides which are unique to single proteins [21,24,25]. A multi-protease approach can also improve the identification of N-Termini and signal peptides for small proteins [26]. Overall, integrating multiple-protease data can increase the number of proteins identified [27,28], the number of identified post-translational modifications detected [24,25,29] and decrease the ambiguity of the protein group list [24].

Lysyl endopeptidase (Lys-C) obtained from *Lysobacter enzymogenesis* is a serine protease involved in cleaving carboxyl terminus of Lys [**PMID:25823410?**]. Like trypsin, the optimum pH range required for its activity is from 7 to 9. A major advantage of Lys-C is its resistance to denaturing agents, including 8 M urea - a chaotrope commonly used to denature proteins *prior* to digestion [**PMID:27123950?**]. Trypsin is less efficient at cleaving Lys than Arg, which could limit the quality of quantitation from tryptic peptides. Hence, to achieve complete protein digestion with minimal missed cleavages, Lys-C is often used simultaneously with trypsin digestion [**PMID:23017020?**].

Alpha-lytic protease (aLP) is also secreted by the soil bacterial *Lysobacter enzymogenesis* [**PMID:3053694?**]. Wild-type aLP (WaLP) and an active site mutant of aLP, M190A (MaLP), have been used to expand proteome coverage [25]. Based on observed peptide sequences from yeast proteome digestion, WaLP showed a specificity for small aliphatic amino acids like alanine, valine, and glycine, but also threonine and serine. MaLP showed specificity for slightly larger amino acids like methionine,

phenylalanine, and surprisingly, a preference for leucine over isoleucine. The specificity of WaLP for threonine enabled the first method for mapping endogenous human SUMO sites [PMID:29079793?].

Glutamyl peptidase I, commonly known as Glu-C or V8 protease, is a serine protease obtained from *Staphyloccous aureus* [PMID:4627743?]. Glu-C cleaves at the C-terminus of glutamate, but also after aspartate [PMID:4627743?,PMID:26748652?].

Peptidyl-Asp metallopeptidase, commonly known as Asp-N, is a metalloprotease obtained from *Pseudomonas fragi* [PMID:2669754?]. Asp-N catalyzes the hydrolysis of peptide bonds at the N-terminal of aspartate residues. The optimum activity of this enzyme occurs at a pH range between 4 and 9.
As with any metalloprotease, chelators like EDTA should be avoided for digestion buffers when using Asp-N. Studies also suggest that Asp-N cleaves at the amino terminus of glutamate when a detergent is present in the proteolysis buffer [PMID:2669754?]. Asp-N often leaves many missed cleavages [PMID:27123950?].

Chymotrypsin or chymotrypsinogen A is a serine protease obtained from porcine or bovine pancreas with an optimum pH range from 7.8 to 8.0 [PMID:3555886?]. It cleaves at the C-terminus of hydrphobic amino acids Phe, Trp, Tyr and barely Met and Leu residues. Since the transmembrane region of membrane proteins commonly lacks tryptic cleavage sites, this enzyme works well with membrane proteins having more hydrophobic residues [PMID:24870543?,PMID:24696503?,PMID:27123950?]. The chymotryptic peptides generated after proteolysis will cover the proteome space orthogonal to that of tryptic peptides both in a quantitative and qualitative manner [PMID:24290761?,PMID:22669647?,PMID:24696503?]

Clostripain, commonly known as Arg-C, is a cysteine protease obtained from *Clostridium histolyticum* [PMID:4332560?]. It hydrolyses mostly the C-terminal Arg residues and sometimes Lys residues, but with less efficiency. The peptides generated are generally longer than that of tryptic peptides. Arg-C is often used with other proteases for improving qualitative proteome data and also for investigating PTMs [PMID:25823410?].

LysargiNase, also known as Ulilysin, is a recently discovered protease belonging to the metalloprotease family. It is a thermophilic protease derived from *Methanosarcina acetivorans* that specifically cleaves at the N-terminus of Lys and Arg residues [PMID:25419962?]. Hence, it enabled discovery of C-terminal peptides that were not observed using trypsin. In addition, it can also cleave modified amino acids such as methylated or dimethylated Arg and Lys [PMID:25419962?].

Peptidyl-Lys metalloendopeptidase, or Lys-N, is an metalloprotease obtained from *Grifola frondosa* [PMID:19195997?]. It cleaves N-terminally of Lys and has an optimal activity at pH 9.0. Unlike trypsin, Lys-N is more resistant to denaturing agents and can be heated up to 70 °C [PMID:25823410?]. Reports suggest that the peptides generated after Lys-N digestion produces more of c-type ions in a ETD-based mass spectrometer [PMID:18425140?]. Hence this can be used for analysing PTMs, identification of C-terminal peptides and also for *de novo* sequencing strategies [PMID:18425140?,PMID:20953479?].

Pepsin A, commonly known as pepsin, is an aspartic protease obtained from bovine or porcine pancreas [PMID:12089768?]. Pepsin was one of several proteins crystalized by John Northrop, who shared the 1946 Nobel prize in chemistry for this work [30/,PMID:19872561?,PMID:19872562?,PMID:17758437?]. Pepsin works at an optimum pH range from 1 to 4 and specifically cleaves Trp, Phe, Tyr and Leu [PMID:25823410?]. Since it possess high enzyme activity and broad specificity at lower pH, it is preferred over other proteases for MS-based disulphide mapping [PMID:12476442?,PMID:24980484?]. Pepsin is also used extensively for structural

mass spectrometry studies with hydrogen-deuterium exchange (HDX) because the rate of back exchange of the amide deuteron is minimized at low pH [31,32].

Proteinase K was first isolated from the mold *Tritirachium album* Limber [PMID:4373242?]. The epithet 'K' is derived from its ability to efficiently hydrolyse keratin [PMID:4373242?]. It is a member of the subtilisin family of proteases and is relatively unspecific with a preference for proteolysis at hydrophobic and aromatic amino acid residues [33]. The optimal enzyme activity is between pH 7.5 and 12. Proteinase K is used at low concentrations for limited proteolysis (LiP) and the detection of protein structural changes in the eponymous technique LiP-MS [PMID:29072706?].

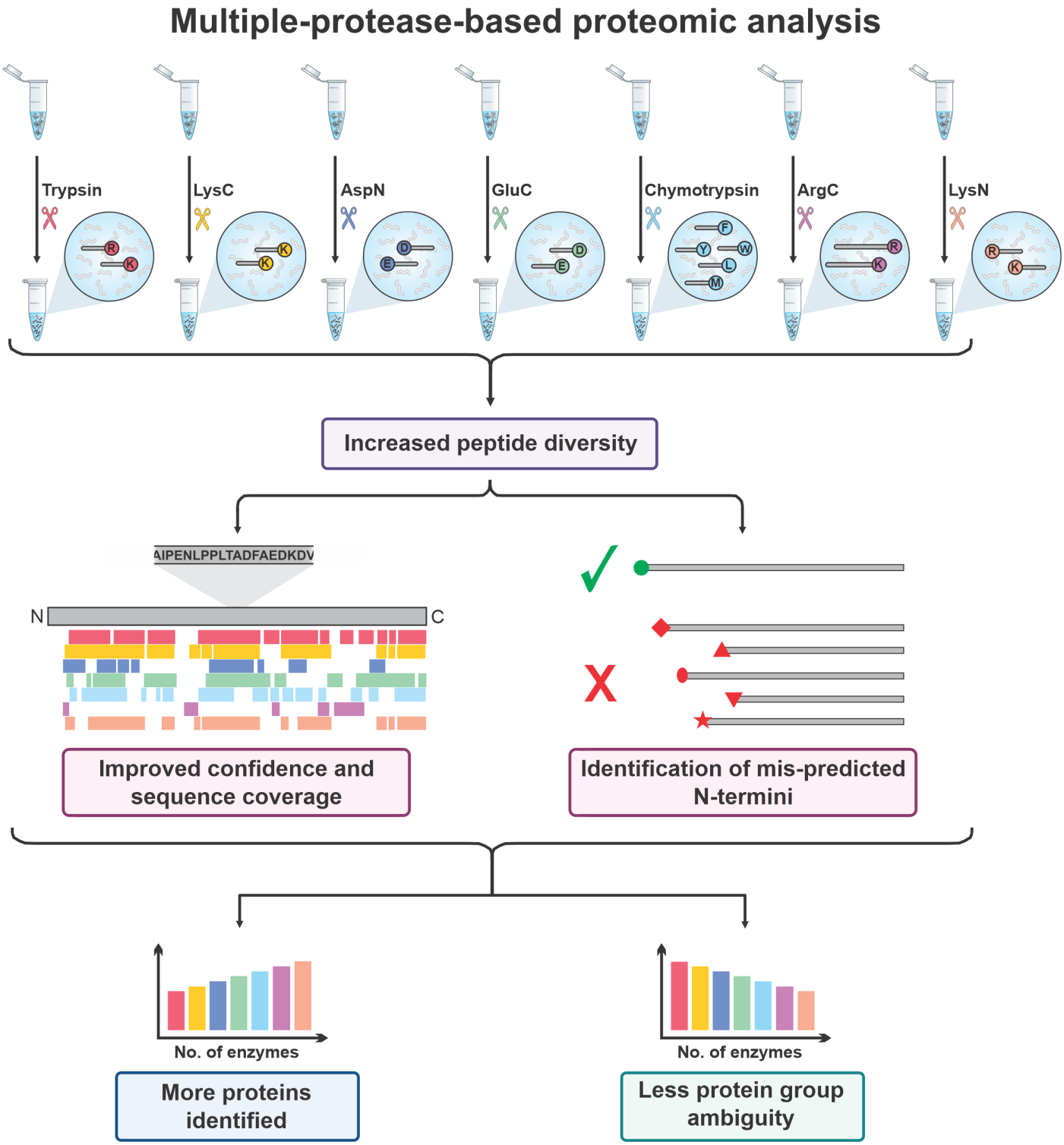

Figure 1: Multiple protease proteolysis improves protein inference The use of other proteases beyond Trypsin such as Lysyl endopeptidase (Lys-C), Peptidyl-Asp metallopeptidase (Asp-N), Glutamyl peptidase I, (Glu-C), Chymotrypsin, Clostripain (Arg-C) or Peptidyl-Lys metalloendopeptidase (Lys-N) can generate a greater diversity of peptides. This improves protein sequence coverage and allows for the correct identification of their N-termini. Increasing the number

of complimentary enzymes used will increase the number of proteins identified by single peptides and decreases the ambiguity of the assignment of protein groups. Therefore, this will allow more protein isoforms and post-translational modifications to be identified than using Trypsin alone.

Peptide Quantification

Discussion of methods to isotopically label peptides or proteins that enable quantification

1. SILAC/SILAM
2. dimethyl labeling
3. Isobaric tags

Peptide labeling with isobaric tags

The isobaric tag labeling-based quantitation uses derivatization of every peptide sample with a different isotope from a set of isobaric mass tags. This is followed by pooling the labelled samples, which undergo MS analysis simultaneously. As the isobaric tags are used, peptides labeled with these tags give a single MS peak with the same precursor m/z value in an MS1 scan and identical retention time of liquid chromatography analysis. The modified parent ions undergo fragmentation during MS/MS analysis generating two kinds of fragment ions: (a) reporter ions and (b) peptide fragment ions. The reporter ions' relative intensity is directly proportional to the amount of peptide in each of the starting samples that were mixed. The fragment ion peaks with higher m/z values correspond to amino acid sequences of peptides and are used for identifying peptides, from which proteins can be inferred. Since it is possible to label most tryptic peptide with an isobaric mass tag at least at the n-termini, numerous peptides from the same protein can be detected, thus leading to an increase in the confidence in both protein identification and quantification [**PMID:25337643?**]. All isobaric tags have a common structural theme consisting of 1) an amine-reactive groups (usually triazine ester or N-hydroxysuccinimide [NHS] esters) which react with peptide N-termini and ε-amino group of the lysine side chain of peptides, 2) a balancer group, and 3) a reporter ion group.

Because the size of the reporter ions is small and sometimes the mass difference between reporter ions is small, these methods are mostly used with high-resolution mass measurement, not with classical ion traps [**PMID:26584918?**]. There are examples, however, of using isobaric tags with pulsed q dissociation on linear ion traps (LTQs) [**PMID:22397766?**]. Suitable instruments are the Thermo Q-Exactive, Exploris lines, and Tribrid lines, or TOFs such as the TripleTOF or timsTOFs [34,**PMID:30967486?**].

The following are some of the isobaric labeling techniques:

isobaric Tags for Relative and Absolute Quantitation (iTRAQ)

The iTRAQ tagging method covalently labels the peptide N-terminus and side-chain primary amines with tags of different masses through the NHS-ester bond. This is followed by mass spectrometry analysis [**PMID:15385600?**]. Reporter ions for an 8plex iTRAQ are measured at roughly 113, 114, 115, 116, 117, 118, 119, and 121 Thompsons. At the moment, two kinds of iTRAQ reagents are available: 4-plex and 8-plex. Using 4-plex reagents, a maximum of four different biological conditions can be analyzed simultaneously, whereas using 8plex reagents, eight different biological conditions can be analysed [**PMID:20593797?**,**PMID:22594965?**].

iTRAQ hydrazide (iTRAQH)

iTRAQH is an isobaric tagging reagent for the selective labeling and relative quantification of carbonyl (CO) groups in proteins [**PMID:22926130?**]. The reactive CO and oxygen groups which are generated as the byproducts of oxidation of lipids at the time of oxidative stress causes protein carbonylation [**PMID:15775985?**]. iTRAQH is produced from iTRAQ and surplus of hydrazine. This reagent reacts with

peptides which are carbonylated, thus forming a hydrazone group. iTRAQH is a novel method for analyzing carbonylation sites in proteins utilizing an isobaric tag for absolute and relative quantitation iTRAQ derivative, iTRAQH, and the analytical power of linear ion trap instruments (QqLIT). This new strategy seems to be well suited for quantifying carbonylation at large scales because it avoids time-consuming enrichment procedures [PMID:22926130?]. Thus, there is no need for enriching modified peptides before LC-MS/MS analysis.

Tandem Mass Tag (TMT)

TMT labeling is based on a similar principle as that of iTRAQ. In the case of 6-plex-TMT, the masses of reporter groups are roughly 126, 127, 128, 129, 130, and 131 Thompsons [PMID:26584918?]. TMT works best with MS which allow quantitation at MS3-level with higher accuracy (e.g.: Thermo's Fusion Orbitrap instruments) [PMID:25337643?] and its eliminates ratio distortion in isobaric multiplexed quantitative proteomics [PMID:21963607?]. The TMT is widely used for quantitative protein biomarker discovery. In addition, TMT labeling technique helps multiplex sample analysis enabling efficient use of instrument time. TMT labelling also controls for technical variation because after samples are mixed the ratios are locked in, and any sample loss would be equal across channels. A wide range of TMT reagents with different multiplexing capabilities are available, such as TMT zero, TMT duplex, TMT six plex, TMT 10-plex, and TMT 11-plex are available along with the recent addition of TMT 16-plex and now even TMT 18-plex [35]. These TMT reagents have a similar chemical structure, which allows the efficient transition from method development to multiplex quantification [PMID:30967486?].

iodoTMT

IodoTMT reagents are isobaric reagents used for tagging cysteine residues of peptides. The commercially available IodoTMT reagents are iodoTMTzero and iodoTMT 6-plex [PMID:24152285?,PMID:24926564?].

aminoxyTMT Isobaric Mass Tags

Also referred to as glyco-TMTs, these reagents have chemistry similar to iTRAQH. The stable isotope-labeled glyco-TMTs are utilized for quantitating N-linked glycans. They are derived from the original TMT reagents with an addition of carbonyl-reactive groups, which involve either hydrazide or aminoxy chemistry as functional groups. These aminooxy TMTs show a better performance as compared to its iTRAQH counterparts in terms of efficiency of labeling and quantification. The glyco-TMT compounds consist of stable isotopes thus enabling (i) isobaric quantification using MS/MS spectra and (ii) quantification in MS1 spectra using heavy/light pairs. Aminoxy TMT6-128 and TMT6-131 along with the hydrazide TMT2-126 and TMT2-127 reagents can be used for isobaric quantification. In the quantification at MS1 level, the light TMT0 and the heavy TMT6 reagents have a difference in mass of 5.0105 Da which is sufficient to separate the isotopic patterns of all common N-glycans. Glycan quantification based on glyco-TMTs generates more accurate quantification in MS1 spectra over a broad dynamic range. Intact proteins or their digests obtained from biological samples are treated with PNGase F/A glycosidases to release the N-linked glycans during the process of labeling using aminoxyTMT reagents. The free glycans are then purified and labeled with the aminoxyTMT reagent at the reducing end. The labeled glycans from individual samples are subsequently pooled and then undergo analysis in MS for identification of glycoforms in the sample and quantification of relative abundance of reporter ions at MS/MS level [PMID:22455665?].

N,N-Dimethyl leucine (DiLeu)

The N,N-Dimethyl leucine, also referred to as DiLeu, is an tandem mass tag reagent which is isobaric and has reporter ions of isotope-encoded dimethylated leucine [PMID:20218596?]. Each incorporated

label produces a 145.1 Da mass shift. A maximum of four samples can be simultaneously analyzed using DiLeu at a highly reduced cost. MS/MS analysis shows intense reporter ions i.e., dimethylated leucine a1 ions at 115, 116, 117, and 118 m/z. The DiLeu tag labeling efficiency is similar to that of the iTRAQ. Although, DiLeu-labeled peptides offer increased confidence for identification of peptides and more reliable quantification as they undergo better fragmentation thus generating higher reporter ion intensities [PMID:20218596?].

Deuterium isobaric Amine Reactive Tag (DiART)

DiART is an isobaric tagging method used in quantitative proteomics [PMID:22404494?,PMID:20715779?]. The reporter group in DiART tags is a N,N′-dimethyl leucine reporter group with a mass to charge range of 114–119. DiART reagents can a label a maximum of six samples and further analyzed by MS. The isotope purity of DiART reagents is very high hence correction of isotopic impurities is not needed at the time of data analysis [PMID:20019052?]. The performances of DiART including the mechanism of fragmentation, the number of proteins identified and the quantification accuracy are similar to iTRAQ. Irrespective of the sequence of the peptide, reporter ions of high-intensity are produced by DiART tags in comparison to those with iTRAQ and thus, DiART labeling can be used to quantify more peptides as well as those with lower abundance, and with reliable results [PMID:22404494?]. DiART serves as a cheaper alternative to TMT and iTRAQ while also having a comparable labeling efficiency. It has been observed that these tags are useful in labeling huge protein quantities from cell lysates before TiO2 enrichment in quantitative phosphoproteomics studies [PMID:24129742?].

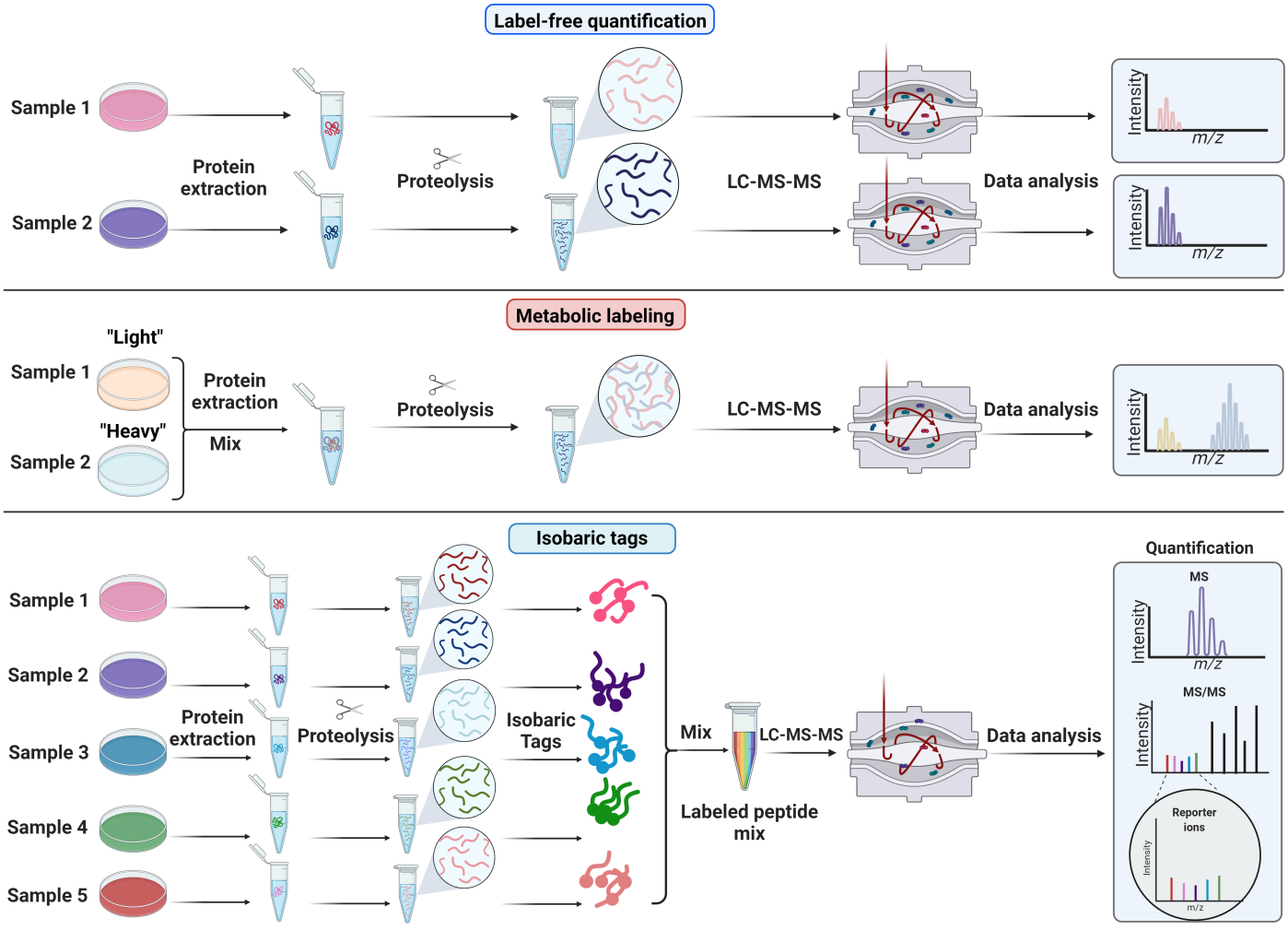

Figure 1: Quantitative strategies commonly used in proteomics. A) Label-free quantitation. Proteins are extracted from samples, enzymatically hydrolyzed into peptides and analyzed by mass spectrometry. Chromatographic peak areas from peptides are compared across samples that are analyzed sequentially. B) Metabolic labelling. Stable isotope labeling with amino acids in cell culture (SILAC) is based on feeding cells stable isotope labeled amino acids (“light” or

“heavy”). Samples grown with heavy or light amino acids are mixed before cell lysis. The relative intensities of the heavy and light peptide are used to compute protein changes between samples. C) Isobaric or chemical labelling. Proteins are isolated separately from samples, enzymatically hydrolyzed into peptides, and then chemically tagged with isobaric stable isotope labels. These isobaric tags produce unique reporter mass-to-charge (m/z) signals that are produced upon fragmentation with MS/MS. Peptide fragment ions are used to identify peptides, and the relative reporter ion signals are used for quantification.

Peptide or Protein Enrichment

Protein enrichment (e.g. for protein protein interactions)

- coIP
- APEX
- bioID
- bioplex

Peptide enrichment

- antibody enrichments of modifications, e.g. lysine acetylation [36].
- TiO2 and Fe enrichment of phosphorylation
- Glycosylation
- SISCAPA

Abundant protein depletion (Blood samples)

The range abundances of proteins in the blood/plasma proteome exceeds 10 orders of magnitude. Due to this wide dynamic range, detection of proteins with medium and low abundance by proteomic analyses is difficult [**PMID:20677825?**]. Identifying protein biomarkers from biological samples such as blood is often obstructed by proteins present at higher concentrations. The removal of these high abundant proteins enables the detection of less abundant and unique proteins. The ability to deplete abundant proteins with specificity, reproducibility, and selectivity is extremely important in proteomic studies [**PMID:16052628?**].

The following are some of the methods used for abundant protein depletion:

Dye-ligand depletion:

This method is used for the depletion of serum albumin based on the interaction between albumin and dyes like Cibacron Blue (CB) through electrostatic force, hydrogen bonding and hydrophobic interactions. The method is relatively low cost, widely available, robust and has high binding capacity. However, it lacks specificity and has varying efficiency [**PMID:11694290?**,**PMID:24168355?**].

Protein-ligand depletion:

This method is used for depletion of immunoglobulins (Ig) based on the interaction between the Fragment crystallizable (Fc) region of these Igs [**PMID:2473373?**] and cell wall protein A, G or A/G of Staphylococcus aureus and Streptococcus spp [**PMID:2938951?**,**PMID:10805799?**]. It is highly selective and has high yield and purity. However, non-specific binding may occur due to co-absorption of other proteins [**PMID:31617391?**].

Immunodepletion:

This method is used for depletion of proteins having high abundance in plasma or serum on the basis of the specific interaction of these proteins with their respective antibodies (antigen-antibody interaction) [PMID:27896769?]. It has high specificity and commercial kits are also readily available, but it is expensive, has limited sample loading and can result in non-specific binding [PMID:31617391?].

Combinatorial peptide ligand library:

This method is used for partial depletion of major proteins i.e., those with high abundance and for relative enrichment of lower and medium abundant proteins [PMID:18451796?]. It is based on the interaction with an array of ligands which are essentially peptides of 6 amino acids in length. It is also used for normalization of the global protein abundance [PMID:25384740?]. However, the drawbacks include non-specific binding as well as loss of proteins due to incomplete elution or inefficient binding [PMID:31617391?].

Precipitation:

This method of abundant protein depletion works by altering the solubility of proteins using a chemical reagent including inorganic salt solution [PMID:21963274?], organic solvents [PMID:25083595?], non-ionic polymer [PMID:27832179?] and reducing agents [PMID:19454248?]. It is extremely simple and cost-effective. However, it is less specific with a risk of protein loss, difficulty in protein resolubilization as well as time consuming [PMID:31617391?].

New technologies:

Newer methods of highly abundant protein depletion are based on the interaction between polymers such as bacterial cellulose nanofibers [PMID:30219335?], cryogels [PMID:30999704?,PMID:23668981?] and nanomaterials [37]. These techniques are highly specific, relatively cheap, and very stable. They can also be reused since they have larger binding capacity and less cross-reactivity [PMID:31617391?].

Peptide Purification and Fractionation

Peptide purification methods

Solid phase extraction (SPE)

Solid phase extraction (SPE) is a common MS-based proteomics technique employed in the sample preparation. In this method, compound isolation is based on chemical and physical properties, which determines the distribution of compounds between a mobile phase (liquid) and a stationary phase (solid). After the molecules bind, washing of the bound compounds is performed and then molecules are made to elute from the stationary phase after replacing the mobile phase with the elution buffer. The material used for SPE is usually discarded after every sample and no gradient is applied for elution (single-step procedure of elution) [PMID:14697044?]. Thus, using SPE only a specific analyte group gets separated, which depends on the stationary phase. Hence, SPE is primarily used for sample clean-up and for reducing complexity of the sample. For MS-based proteomic analysis, it is largely used to get rid of salts and other contaminants that might lead to ion suppression. The major drawback of this technique is that with SPE only a small fraction of the sample is examined because not all compounds are captured, but only those with binding capabilities same as that of the sorbent. The material for SPE is available in various types, including (micro-) columns, cartridges, plates, micropipette tips, and functionalized magnetic beads (MBs) [PMID:20606758?,PMID:20099258?]. Reversed-phase is the most widely used material for SPE in proteomic studies for the proteins and peptide fractionation and rarely, ion-exchange material. For the separation of glycosylated proteins and peptides, the preferred material is normal phase such as HILIC [[PMID:22665312?]; 20536156]. SPE materials which are less commonly used are silica- or polystyrene-based ones [[PMID:17625912?]; 15317408]. The other types of SPE methods are IEX, metal chelation, and affinity-based [PMID:25692071?].

The basic idea behind the choice of binding and wash versus elution solutions for SPE is that that the binding and wash solutions should favor the interaction between the analytes of interest and the solid phase, whereas the elution solution should favor the interaction of the analyte with the liquid phase. For example, with reversed phase SPE, the solid phase is C18 or some other hydrophic chemistry. Binding of peptides to this solid phase is based on the hydrophobicity of peptides, mostly due to their peptide backbone, but also due to the presence of amino acid side chains like leucine and phenylalanine. To encourage peptides to ‘like’ the stationary phase more than the liquid phase, the peptides are loaded in aqueous solution. This will enable washing of the hydrophilic contaminants like salts, small polar buffer molecules, and polar denaturants like urea. After washing the bound peptides, they can be eluted by switching the liquid phase to something hydrophobic, which allows the peptides to partition more into the liquid phase and elute from the solid phase.

Specific Types of peptide purification

1. Reverse phase including tips and cartridges
2. stage tips
3. in stage tip (iST)
4. SP2, SP3
5. s traps

Peptide fractionation methods

The number of peptides produced from proteolysis of the whole proteome is immense. Thus, after peptides are cleaned from interferences, they are often fractionated into subsets to enable increased

proteome coverage. The characterization of the whole proteome is expected from higher order organisms, and with rising interest in post-translational modifications, an elaborate coverage of protein sequence is required. There are different methods for peptide fractionation as follows:

Ion-exchange chromatography (IEC)

This method involves the separation based on contrasting electric charge [PMID:27868236?]. In this approach, the mechanism of analyte retention is based on the principle of electrostatic attraction between the sample and the stationary phase functional groups (FGs), having opposite charges. IEC is classified into two types: cation-exchange and anion-exchange chromatography. In cation-exchange chromatography, at an acidic pH, the negatively charged functional groups such as sulfates are attracted to positively charged peptides, whereas, in anion-exchange chromatography, positively charged FGs such as quaternary ammoniums are attracted to peptides with negative charge at an alkaline pH. These techniques are further classified into: strong (cation [SCX] and anion [SAX] exchange), and weak exchangers (cation [WCX] and anion [WAX] exchange), based on the type of FG attached [PMID:35777803?]. These functional groups are most commonly supported in resins made up of silica and synthetic polymers, however, some inorganic materials are sometimes used [PMID:27868236?]. In the IEC method, peptide elution is performed using a mobile phase with higher ionic strength, to ensure peptide partition into the liquid phase. SCX along with a salt gradient/plug is a routinely used proteomics technique. In the SCX method, peptides are resolved according to their net charge, in which the peptide with the lowest positive charge is eluted first. Increasing the salt concentration decreases the peptide retention time due to competition with the electrostatic interactions between the peptides and the solid phase. However, SCX resolution is limited compared to reversed phase chromatography and will thus limit the suitability of this technique for complex mixtures [PMID:15672457?].

Reversed-phase chromatography (RPLC)

Reversed-phase chromatography is the most widely commonly used chromatographic technique which separates molecules in solution having neutral pH based on their hydrophobicity. The separation occurs on the basis of the partition coefficient of analytes between the mobile phase and the hydrophobic stationary phase. Highly polar peptides elute before the ones having less polarity because of the strong interaction with the hydrophobic functional groups forming a layer similar to a liquid around the silica resin [PMID:20973639?]. RPLC has been widely used in separation of peptides because of its compatibility with gradient elution and aqueous samples and its retention mechanism, which modulates separation owing to changes in the properties like pH, additives and organic modifier [PMID:20031138?]. Numerous factors influence the capacity of chromatographic peaks, such as temperature, column length, stationary phase, particle size, mobile-phase ion-pairing reagent, mobile-phase modifier and gradient slope[PMID:16224963?]. Usually online RPLC is done at acidic pH to ensure peptide ionization, but it can be paired with offline high pH RPLC and multiple fraction concatenation to produce orthogonal separation due to altered ionization of amino acids changing peptide hydrophobicity [PMID:22462785?].

Hydrophilic interaction liquid chromatography (HILIC)

HILIC is similar in its principle to normal-phase chromatography. It is used for the separation of hydrophilic peptides and polar analytes [PMID:21879300?]. This separation is achieved by a stationary phase that is hydrophilic in nature, for example: cyano-, diol-, amino- bonded phases [PMID:18428181?], and an organic and hydrophobic mobile phase [PMID:18264818?]. The elution of bonded peptides occurs by increasing the mobile phase polarity in a reversed elution order as compared to RPLC [[PMID:18264818?]; 15459207]. Thus, the peptides with less polarity elute before the more polar peptides. HILIC can also be used for enrichment and targeted proteomic analysis of

PTMs, such as glycosylation, N-acetylation and phosphorylation, which increase the polarity of peptides and therefore also their retention on HILIC [**PMID:20973639?**].

Isoelectric focusing (IEF)

IEF is a type of high-resolution (HR) technique of electrophoresis used for the separation as well as concentration of peptides that are amphoteric in nature on the basis of their isoelectric point (pI) using a solution without buffer consisting of either carrier ampholytes or a gel with immobilized pH gradient (IPG). After IEF separation, the separated amphoteric peptides in the liquid phase are recovered for further analysis by RPLC-MS/MS [**PMID:16849286?**]. Along with being a technique with improved resolution and capacity, for separation of peptides, IEF provides with additional information on physicochemical properties of the peptides, for example: peptide iso electric point (pI) which acts as a tool for validation and filtration for identifying MS/MS peptide sequence during the step of database search [**PMID:18851748?**]. The IEF system is not only used for increasing the coverage of proteome but also in quantitative label-free [**PMID:17708596?**] and stable isobaric labeling experiments [**PMID:18851748?**]. IEF and gel-based separations have fallen out of favor in the last decade due to improvements in liquid chromatography.

Electrostatic repulsion-hydrophilic interaction chromatography (ERLIC)

ERLIC is a method based on use of a weak anion exchange column operated at low pH with high organic solvent enabling isocratic elution [38]. Acidic peptides are retained by electrostatic interaction, basic and neutral peptides are retained through hydrophilic interaction made favorable by high organic solvent. This improves retention of acidic peptides and reduces retention of basic peptides compared to normal HILIC [39].

Types of Mass Spectrometers used for Proteomics

1. QQQ
2. Q-TOF
3. Q-Orbitrap
4. LTQ-Orbitrap
5. TOF/TOF
6. FT-ICR
7. types of ion mobility

- SLIM
- FAIMS
- traveling wave
- tims

Peptide Ionization

Until the early 1990s, peptides analysis by mass spectrometry was challenging. Hard ionization techniques in use at the time, like fast atom bombardment, were not directly applicable to peptides without destroying or breaking them. The soft ionization tehniques however, revolutionized the proteomics field and it became possible to routinely ionize and analyze peptides using MALDI and ESI techniques at high-throughput scale. These two techniques were so impactful that the 2002 Nobel Prize in Chemistry was co-awarded to John Fenn (ESI) and Koichi Tanaka (MALDI) “for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules” [40].

MALDI

The term, Matrix-assisted LASER desorption ionization (MALDI), was coined by Hillenkamp and Karas in 1985[41]. Karas and Hillenkamp discovered the MALDI technique first, although a similar ionization method was shown by Koichi Tanaka in 1988 [2]. A few months later, Karas and Hillenkamp also demonstrated MALDI applied to protein ionization [42]. It also created a controversy that the widely used method of MALDI from these two people had been overlooked, and the Nobel prize was awarded to Tanaka, whose system was rarely used[43].

MALDI first requires the peptide sample to be co-crystallized with a matrix molecule, which is usually a volatile, low molecular-weight, organic aromatic compound. Some examples of such compounds are cyno-hydroxycinnamic acid, dihyrobenzic acid, sinapinic acid, alpha-hydroxycinnamic acid, ferulic acid etc [44/]. Subsequently, the analyte is placed in a vacuum chamber in which it is irradiated with a LASER, usually at 337nm [45]. This laser energy is absorbed by the matrix, which then transfers that energy along with its free protons to the co-crystalized peptides without significantly breaking them. The matrix and co-crystallized sample generate plumes, and the volatile matrix imparts its protons to the peptides as it gets ionized first. The weak acidic conditions used as well as the acidic nature of the matrix allows easy exchange of protons for the peptides to get ionized and fly under the electrical field in the mass spectrometer. These ionized peptides generally form the metastable ions, most of them will fragment quickly [46]. However, it can take several milliseconds and the mass spectrometry analysis can be performed before this time. Peptides ionized by MALDI almost always take up a single charge and thus observed and detected as [M+H]+ species.

MALDI Mechanism

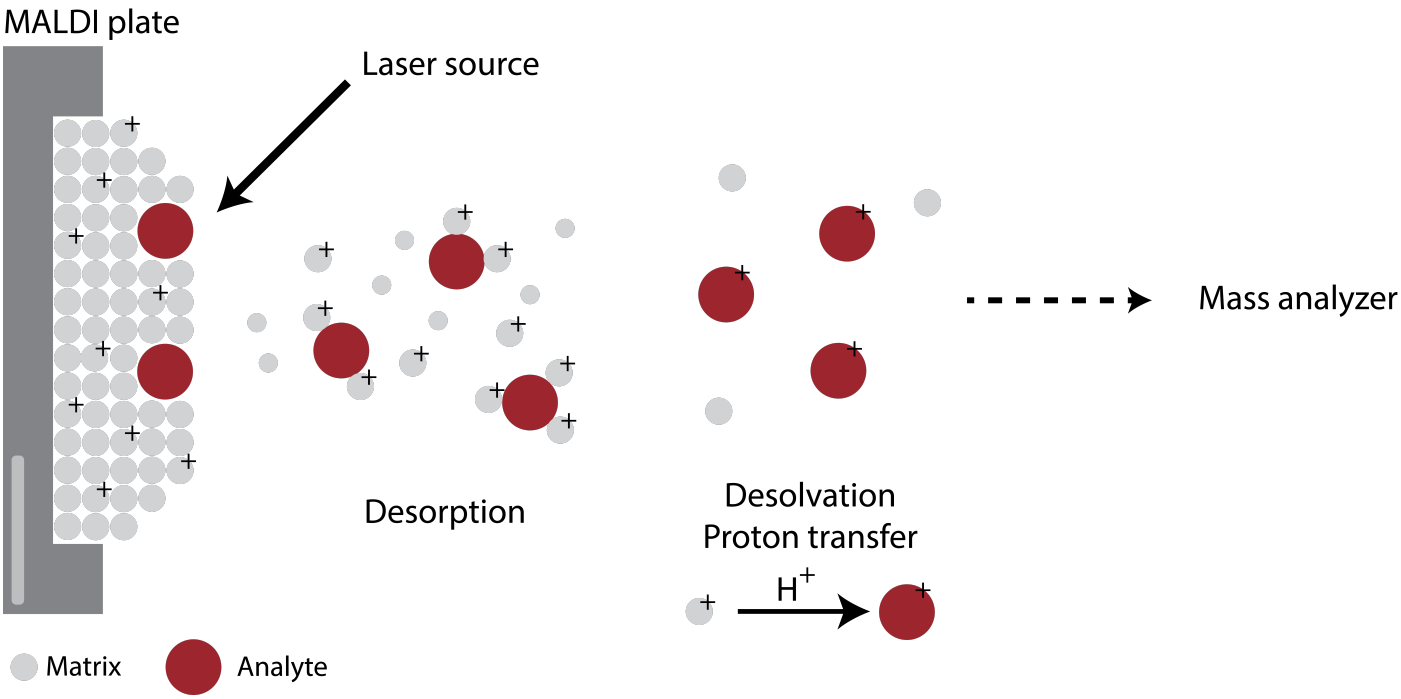

Figure 1: MALDI The analyte-matrix mixture is irradiated by a laser source, leading to ablation. Desorption and proton transfer ionize the analyte molecules that can then be accelerated into a mass spectrometer.

Electrospray Ionization

ESI was first applied to peptides by John Fenn and coworkers in 1989 [1]. Concepts related to electrospray Ionization (ESI) were published at least as early as 1882, when Lord Rayleigh described the number of charges that could assemble on the surface of a droplet [47]. ESI is usually coupled with reverse-phase liquid-chromatography of peptides directly interfaced to a mass spectrometer. A high voltage (~ 2 kV) is applied between the spray needle and the mass spectrometer. As solvent exits the needle, it forms droplets that take on charge at the surface, and through a debated mechanism, those charges are imparted to peptide ions. The liquid phase is generally kept acidic to help impart protons easily to the analytes.

Tryptic peptides ionized by ESI usually carry one charge one the side chain of their c-terminal residue (Arg or Lys) and one charge at their n-terminal amine. Peptides can have more than one charge if they have a longer peptide backbone, have histidine residues, or have missed cleavages leaving extra Arg and Lys. In most cases, peptides ionized by ESI are observed at more than one charge state. Evidence suggests that the distribution of peptide charge states can be maniupulated through chemical additives [**PMID:22610994?**].

Electrospray Mechanism

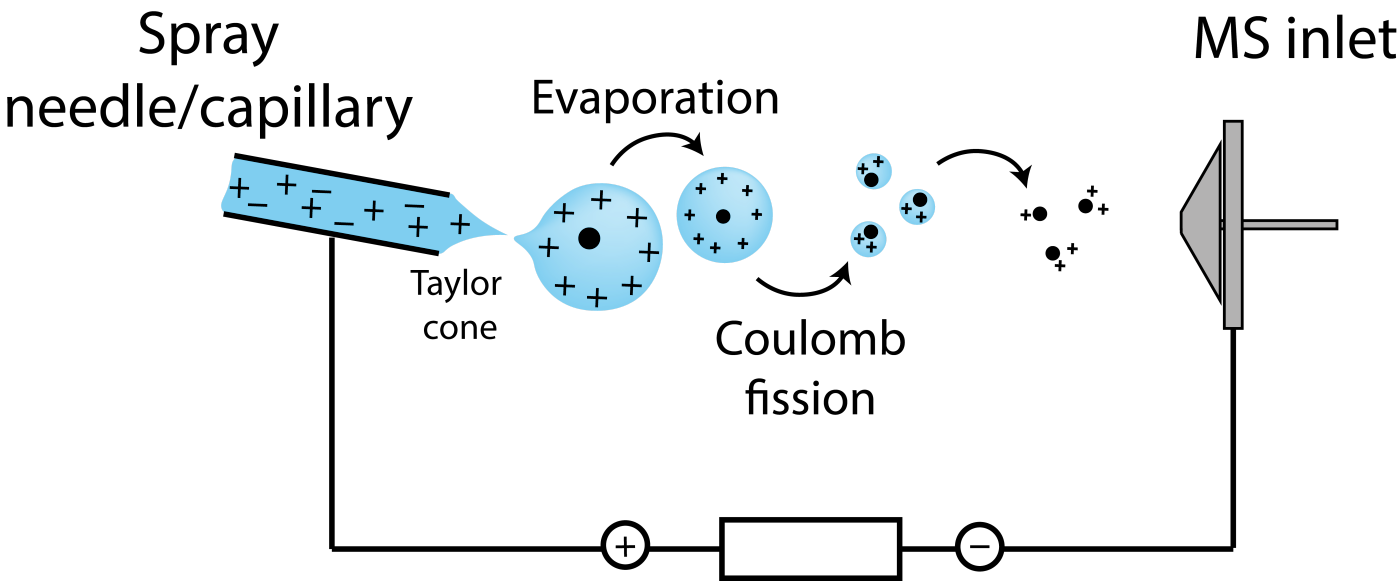

Figure 1: Electrospray Ionization Charged droplets are formed, their size is reduced due to evaporation until charge repulsion leads to Coulomb fission and results in charged analyte molecules.

The main goal of ESI is the production of gas-phase ions from electrolyte ions in solution. During the process of ionization, the solution emerging from the electrospray needle or capillary is distorted into a Taylor cone and charged droplets are formed. The charged droplets subsequently decrease in size due to solvent evaporation. As the droplets shrink, the charge density and Coulombic repulsion increase. This process destabilizes the droplets until the repulsion between the charges is higher than the surface tension and they fission (Coulomb explosion) [PMID:19551695?] [PMID:23134552?]. Typical bottom-up proteomics experiments make use of acidic analyte solutions which leads to the formation of positively charged analyte molecules due to an excess presence of protons.

Data Acquisition

Hybrid mass spectrometers used for modern proteome analysis offer the flexibility to collect data in many different ways. Data acquisition strategies differ in the sequence of precursor scans and fragment ion scans, and in how analytes are chosen for MS/MS. Constant innovation to develop better data collection methods improves our view of the proteome, but many method options may confuse newcomers. This section provides an overview of the general classes of data collection methods.

Data acquisition strategies for proteomics fall into one of two groups.

1. Data dependent acquisition (DDA), in which the exact scan sequence in each analysis depends on the data that the mass spectrometer observes.
2. Data independent acquisition (DIA), in which the exact scan sequence in each analysis DOES NOT depend on the data; the collected scans are the same whether you inject yeast peptides, human peptides, or a solvent blank.

DDA and DIA can both be futher subdivided in to targeted and untargeted methods.

DDA

In most cases, the peptide masses that will be observed are not known before doing the experiment. Data collection methods must account for this. DDA was invented in the early 1990s, which enabled collecting MS/MS spectra for observed peptides as they eluting from the LC column [48,49,**PMID:24203425?**].

Untargeted DDA

A common method currently used in modern proteomics is untargeted DDA. The MS collects precursor (MS1) scans iteratively until precursor mass envelopes meeting certain criteria are detected. Criteria for selection are usually specific charge states and a minimum signal intensity. When those ions meet these criteria, the MS selects those masses for fragmentation.

Because ions are selected as they are observed, repeated DDA of the same sample will produce a different set of identifications. This stochasticity is the main drawback of DDA. To ameliorate this issue, often strategies are used to transfer identifications between multiple sample analyses. This transfer of IDs across runs is known as “match between runs”, which was originally made famous by the processing software MaxQaunt [50,51]. There are several other similar tools and strategies, including the accurate mass and time approach [52], Q-MEND [53], IDEAL-Q [54] and superHIRN [55]. More recent work has introducted statistical assessment of MBR methods using a two-proteome model [56]. Statistically controlled MBR is currently available in the IonQuant tool [57].

Because DDA is required for quantification of proteins using isobaric tags like TMT, this stochasiticity of DDA limits the ability to compare quantities across batches. For example, if you have 3o samples, you can use two sets of the 16-plex kit to label 15 samples in each set with one channel labeled by a pooled sample to enable comparison across the groups. When you collect DDA data from each of those sets, each set will have MS/MS data from an overlapping but different set of peptides. If one set has MS/MS from a peptide but the other set does not, then that peptide cannot be quantified in the whole sample group. This limits the number of quantified proteins in large TMT experiments with multiple batches.

Targeted DDA

Targeted DDA is not common in modern proteomics. In targeted DDA, in addition to general criteria like a minimum intensity and a certain charge state, the mass spectrometer looks for specific masses. These masses might be previously observed signals that were previously missed by MS/MS [58,59]. In these studies, the sample is first analyzed by LC-MS to detect precursor ion features with some software, and then subsequent analyses target those masses for fragmentation with inclusion lists until they are all fragmeted. This was shown to increase proteome converage.

DDA methods for modifications

There are also DDA methods that look for specific fragment or neutral loss ions in the resulting spectra. For example, when linear ion traps were the main proteomics workhorses, CID analysis of phosphopeptides would result in predominantly neutral loss of the phosphate with limited sequence ion information. To gain sequence ions in these experiemnts, instruments could be set to isolate a loss of 98 Thompsons for MS3 [60,61]. The newer collisional dissociation technique HCD significantly improves the detection of peptide fragments with the phosphorylation intact on fragment ions, and thus, this neutral loss scanning technique is no longer common.

A similar strategy was introduced for N-linked glycopeptides [62]. Collisional dissociation of glycosylated peptides produces oxonium ions at 204.09 (HexNAc) or 366.14 (HexHexNAc). In oxonium ions from the glycosylation were detected in the top 20 fragments of the HCD spectra, then an ETD scan was triggered. This ETD scan provides information about the peptide sequence, while the original HCD scan provides glycan structure information.

DIA

The simplest method to operate a mass spectrometer is to have predefined scans that are collected for each sample analysis. This is DIA; the scans that are collected do not depend on the data that the instrument observes. Thus, the scan sequence is the same every time. Although simple in terms of data collection, when the scan sequence includes MS/MS, sophisticated software is required to analyze the data. Like DDA, DIA can also be either targeted or untargeted [63]: The two targeted DIA methods are selected reaction monitoring (SRM) or multiple reaction monitoring (MRM), and untargeted DIA (uDIA) is often referred to simply as “DIA” or SWATH (Figure 4).

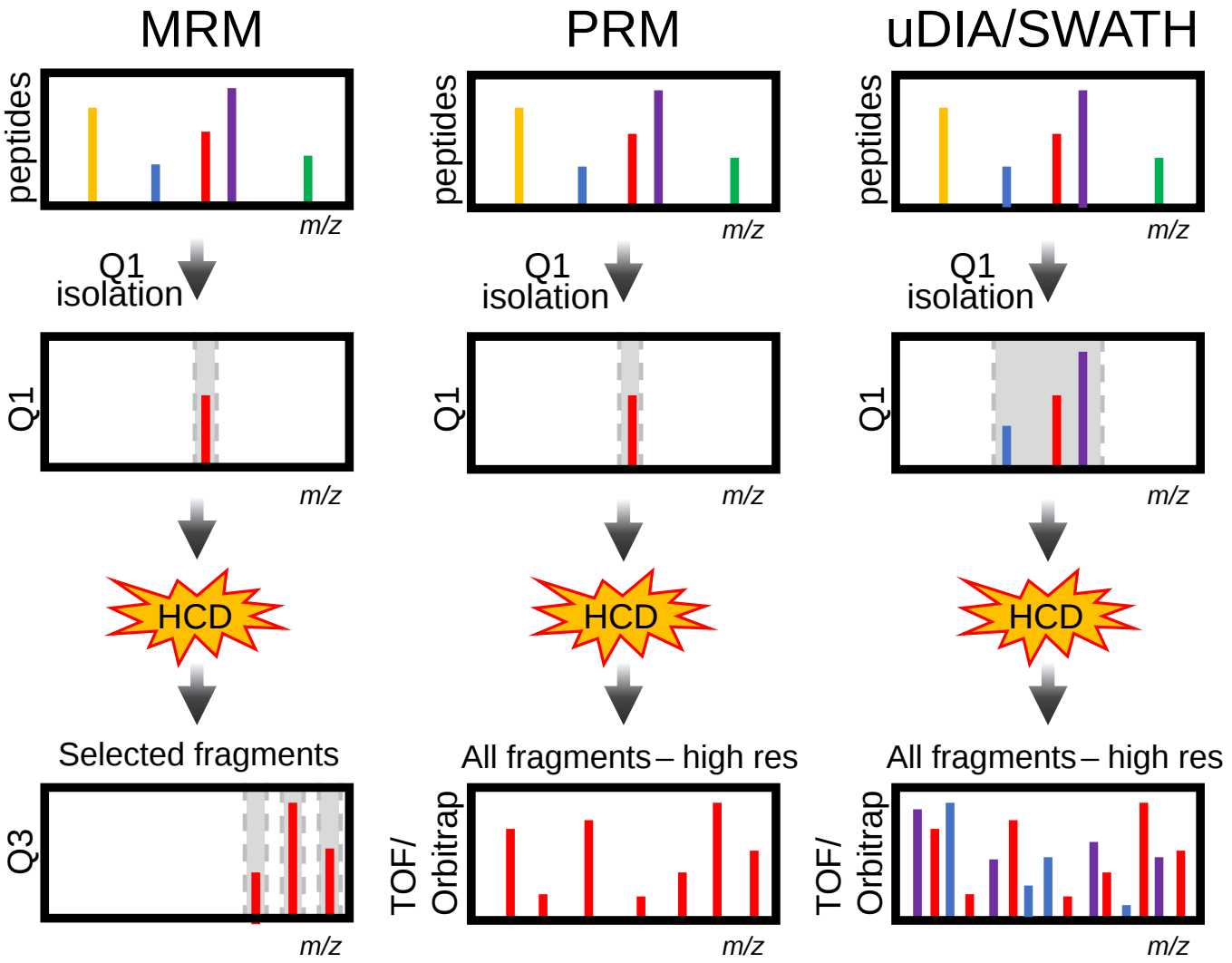

Figure 4: Types of DIA. A) SRM/MRM. Peptides are ionized by ESI and although there are many peptides entering the mass spectrometer at any time, the first quadruople (Q1) isolates one mass, which is then fragmented by HCD. Fragment masses from the peptide are then selected in the third quadruople (Q3). This leads to very low noise and high sensitivity. B) PRM. Like MRM, peptides are selected in the first quadruople, but this analysis is done on a high resolution instrument like an Orbitrap or TOF. Selectivity is gained by exploiting the high mass accuracy and resolution to monitor multiple fragment ions. C) uDIA/SWATH. Like MRM and PRM, peptides are isolated with Q1, but in this case a much wider isolation window is used. This usually results in co-isolation of many peptides simultaneously. Fragments from many peptides are measured with high resolution and high mass accuracy. Special software is used to get peptide identities and quantities from the fragment ions.

Targeted DIA

The first type of targeted DIA is called SRM or MRM [64]. The popularity of this method in the literature peaked in 2014, with just under 1,500 documents on pubmed that year resulting from a search for “MRM”. In this strategy, the QQQ MS is set so that the first quadrupole selects the precursor mass of the peptide(s) of interest, the second quadrupole fragments the peptide, and the third quadruople monitors the product of specific fragments from that peptide. This strategy is very sensitive and has the benefit of very low noise. The fragments monitored in Q3 are chosen such that it is unlikely these fragments could arise from another peptide. Usually at least a few transitions are monitored for each peptide in order to get multiple measures for that peptide.

An early example of MRM applied to quantify c-reactive protein was in 2004 [65]. Around the same time, SRM was combined with antibody enrichment of peptides from target proteins [66]. This approach was popular for analysis of plasma proteins [67]. These early examples led to many more studies that used QQQ MS instruments to get accurate quantitation of many proteins in one injection

[68,69]. Scheduling MRM measurement when chromatography is stable additionally enabled better utilization of instrument duty cycle and therefore monitoring of more peptides per injection [70]. Efforts even developed libraries of transitions that allow quantification of any protein in model organisms [71].

Another similar targeted DIA method is called parallel reaction monitoring (PRM) [72]. Instead of using a QQQ instrument, PRM uses a hybrid MS with a quadrupole and a high resolution mass analyzer, such as an Q-TOF or Q-Exactive. The idea is that instead of monitoring specific fragments in Q3, the high mass accuracy can be used to filter peptide fragments for high selectivity and accurate quantification. Studies have found that PRM and MRM/SRM have comparable dynamic range and linearity [73].

Untargeted DIA

There were many implementations of uDIA over the years, starting in 2003 by Purvine et al from the Goodlett lab [74]. In this first work they demonstrated uDIA using a Q-TOF with in source fragmentation, and showed that extracted ion chromatograms of precursor and fragment ions matched in shape suggesting that this could be used to identify and quantify peptides. The following year, Venable et al from the Yates lab introduced uDIA with an ion trap [75]. Subsequent methods include MSE [76], PAcIFIC [77], all ions fragmentation (AIF) [78]. Computational methods were also developed to automate interpretation of this data, such as DeMux [78], XDIA [79], and ETISEQ [80].

The paper that is often cited for uDIA that led to widespread adoption was by Gillet et al. from the Aebersold group in 2012 [81]. In this paper they branded the idea as SWATH. Widespread adoption may have been facilitated by the co-marketing of this idea by ABSciex as a proteomics solution on their new 5600 Q-TOF (called “tripleTOF” despite containing only one TOF, likely a portmanteau of “triple quadrupole” and “Q-TOF”). Importantly, in the Gillet et al. paper the authors described a computational method to extract information from SWATH where peptides of interest were queried against the data. They also demonstrated the application of SWATH to measure proteomic changes that happen in diauxic shift, and showed that SWATH can reveal modified peptides, in this case a methionine oxidation.

There are also many papers describing uDIA with orbitraps. One early example described combining random isolation windows together and then demultiplexing the chimeric spectra [82]. In another landmark paper, over 6,000 proteins were identified from mouse tissue by at least 2 peptides [83]. In 2018, the new model orbitrap at that time (HF-X) enabled identification of nearly 6,000 human proteins in only 30 minutes. Currently orbitraps have all but replaced the Sciex Q-TOFs for DIA data collection.

A new direction in uDIA is the addition of ion separation by ion mobility. This has appeared in two forms. On the timsTOF, diaPASEF makes use of the trapped ion mobility to increase speed and sensitivity of analysis [84]. On the orbitrap, the combination of FAIMS and DIA has enabled the identification of over 10,000 proteins from one sample, which is a major milestone [85].

Analysis of Raw Data

The goal of basic data analysis is to convert raw spectral data into identities and quantities of peptides and proteins that can be used for biologically-focused analysis. This step may often include measures of quality control, cross-run data normalization, quantification on different levels (precursor, peptide, protein), protein inference, PTM (post translational modification) localization and also first steps of data analysis, such as statistical hypothesis tests.

In typical bottom-up proteomics experiments, proteins are digested into peptides and further analyzed with LC-MS/MS systems. Peptides can have different PTMs and ionize differently depending on their length and amino acid distributions. Therefore, mass spectrometers often record different charge and modification states of one single peptide. The entity that is recorded on a mass spectrometer is usually referred to as a precursor ion (peptide with its modification and charge state). This precursor ion is fragmented and the precursor or peptide sequences are obtained though spectral matching. The quantity of a precursor is estimated with various methods. The measured precursor quantities are combined to generate a peptide quantity. Peptides are also often combined into a protein group through protein inference, which combines multiple peptide identifications into a single protein identification [**PMID:16009968?**] [86]. Protein inference is still a challenge in bottom-up proteomics.

Due to the inherent differences in the data structures of DDA and DIA measurements, there exist different types of software that can facilitate the steps mentioned above. The existing software for DDA and DIA analysis can be further divided into freeware and non-freeware:

DDA freeware

Name	Publication	Website
MaxQuant	Cox and Mann, 2008[87]	MaxQuant
MSFragger	Kong et al., 2017[88]	MSFragger
Mascot	Perkins et al., 1999[**PMID:10612281?**]	Mascot
MS-GF+	Kim et al., [89]	MS-GF+
X!Tandem	Craig et al., [90,91]	GPMDB

DIA freeware:

Name	Publication	Website
MaxDIA	Cox and Mann, 2008[87]	MaxQuant
Skyline	MacLean et al., 2010[92]	Skyline
DIA-NN	Demichev et al., 2019[93]	DIA-NN

Targeted proteomics freeware:

Name	Publication	Website
Skyline	MacLean et al., 2010[92]	Skyline

DDA non-freeware:

Name	Publication	Website
ProteomeDiscoverer		ProteomeDiscoverer
Mascot	Perkins et al., 1999[PMID:10612281?]	Mascot
Spectromine		Spectromine
PEAKS	Tran et al., 2018[94]	PEAKS

DIA non-freeware:

Name	Publication	Website
Spectronaut	Bruderer et al., 2015[95]	Spectronaut
PEAKS	Tran et al., 2018[94]	PEAKS

Data Summary and Interpretation

Name	Publication	Website
Peptide Shaker	Vaudel *et al.*, 2015[96,97]	PeptideShaker, Peptide Shaker Online

Analysis of DDA data

DDA data analysis either directly uses the vendor proprietary data format directly with a proprietary search engine like Mascot, Sequest (through Proteome Discoverer), Paragon (through Protein Pilot), or it can be processed through one of the many freely available search engines or pipelines, for example, MaxQuant, MSGF+, X!Tandem, Morpheus, MSFragger, and OMSSA. Tables 1 and 4 give weblinks and citations for these software tools. For analysis with freeware, raw data is converted to either text-based MGF (mascot generic format) or into a standard open XML format like mzML [98] [[PMID:20013381?]][99]. The appropriate FASTA file containing proteins predicted from that organism's genome is chosen as a reference database to search the experimental spectra. All search parameters like peptide and fragment mass errors (i.e. MS1 and MS2 tolerances), enzyme specificity, number of missed cleavages, chemical artefacts (fixed modifications) and potential biological modifications (variable/dynamic modifications) are specified before executing the search. The search algorithm scores each query spectrum against its possible peptide matches [100]. A spectrum and its best scoring candidate peptide are called a peptide spectrum match (PSM). The scores reflect a *goodness-of-fit* between an experimental spectrum and a theoretical one and do not necessarily depict the correctness of the peptide assignment.

For evaluating the matches, a decoy database is preferred as a null model for peptide matching. A randomized or reversed version of target database is used as a nonparametric null model. The decoy database can be searched separate from the target database (Kall's method)[101] or it can be combined with the target database before search (Elias and Gygi method)[PMID:17327847?]. Using either separate method or concatenated database search method, an estimate of false hits can be calculated which is used to estimate the false discovery rate (FDR) [102]. The FDR denotes the proportion of false hits in the population accepted as true. For Kall's method: the false hits are estimated to be the number of decoys above a given threshold. It is assumed that the number of decoy hits that pass a threshold are the false hits. A similar number of target population may also be false. Therefore, the FDR is calculated as [103]:

$$FDR = \frac{DecoyPSMs + 1}{TargetPSMs}$$

For Elias and Gygi Method, the target population in which FDR is estimated changes. The target and decoy hits coming from a joint database compete against each other. For any spectrum, either a target or a decoy peptide can be the best hit. It is argued that the joint target-decoy population has decoy hits as confirmed false hits. However, due to the joint database search, the target database may also have equal number of false hits. Thus, the number of false hits is multiplied by two for FDR estimation.

$$FDR = \frac{2 * DecoyPSMs}{Target + DecoyPSMs}$$

Strategies for analysis of DIA data

Targeted proteomics data analysis

Quality control

Quality control should be a central aspect of any mass spectrometry-based study to ensure reproducibility of generated results. There are two types of quality controls that can be conducted for any kind of mass spectrometry experiment. The first one is focused on monitoring the perfomance of the instruments themselves (e.g. HPLC and mass spectrometer), whereas the second one is focused on your experiments. For further reading, we recommend to take a look at issue 11 on quality control published in the journal *Proteomics* in 2011 [104], especially the review by Köcher *et al.* [105], as well as the review published by Bittremieux *et al.* in 2017 [106].

Instrument Performance

It is generally advisable to monitor instrument performance regularly. Instrument calibrations in regular intervals help ensure that performance is maintained. Often basic calibration and sensitivity can be checked by direct infusion of a standard. During the calibration you can check injection times (for ion trap instruments) and intensity of the ions in the calibration mix.

After ensuring good calibration and signal with the simple calibration mixture, it is advisable to analyze complex samples, such as tryptic digests of whole-cell lysates (e.g. HeLa cells, HEK cells, yeast, etc.) or tryptic digests of purified proteins. The additional check with a complex sample ensures all aspects of the system are working together correctly, especially the liquid chromatography and emitter. These digests should be analyzed after every instrument calibration and periodically between samples when acquiring more extensive batches. Data measured from tryptic digests should be analyzed by the software of your choice and the numbers of identified peptide precursors and proteins can be compared with previous controls for consistency.

Another strategy is to analyze digested purified proteins, which easily enable discovery of retention time shifts and mass accuracy problems. In case you are working with a Thermo mass spectrometer, you can open the acquired .raw file directly either in FreeStyle or in Qual Browser and look for specific m/z values of your peptides. Looking at the intensity of the extracted peaks will help identify sensitivity fluctuations.

Carry-over between different measurments can be identified from blank measurements which are subsequently analyzed with your search software of choice. Blank measurements can be injections of

different buffers, water or the starting conditions of your liquid chromatography. In case of increased detection of carry-over, injections with trifluoroethanol can be performed.

Another factor to take into consideration is the stability of your electrospray. Electrospray stability tends to worsen over time as columns wear, as well as when measuring samples with residual contaminants, such as salts or detergents. You will notice spray instabilities either in the total ion chromatogram (TIC) as thin spikes with short periods of no measured signal or if you install cameras at your ESI source. Suboptimal spray conditions will usually result in droplets forming on the emitter, being released into the mass spectrometer (also referred to as “spitting”). Real-time quality control software (listed in the table below) can help you identify instrument issues right away.

Data Quality Control

Apart from instrument performance, any kind of data analysis should have proper quality control in place to identify problematic measurements and to exclude them if necessary. It is recommended to develop a standardized system for data quality control early on and to keep this consistent over time. Adding indexed retention time (iRT) peptides can help identify and correct gradient and retention time inconsistencies between samples at the data analysis stage. Decoy searches help monitor and control the false-discovery rate. Inlcuding common contaminants, such as keratins, in the FASTA files used for searches can help identify sample preparation issues. Other parameters to check in your analysis are the consistency of the number of peptide-spectrum matches, identified peptides and proteins over all samples of your study, as well as your coefficients of variation between your replicates. Before and after data normalization (if normalization is performed) it is good to compare the median intensities of all measurments to identify potential measurement or normalization issues. Precursor charge distributions, missed cleavage numbers, peak width, as well as the number of points per peak are additional parameters that can be checked. In case you are analyzing different conditions, you can perform hierarchical clustering or a principal component analysis to check if your samples cluster as expected.

Quality Control Software

Raw file and real-time analysis

Name	Supported instrument vendors	Website/Download	publication	Note
QuiC	Thermo Scientific, AB SCIEX, Agilent, Bruker, Waters	QuiC		requires Biognosys iRT peptides
Alpha Pept	Thermo Scientific, Bruker	AlphaPept	[107]	
RawMeat 2.1	Thermo Scientific	RawMeat		
rawDiag	Thermo Scientific	rawDiag	[108]	
rawrr	Thermo Scientific	rawrr	[109]	
rawBeans	Thermo or mzML	rawBeans	[110]	
SIMPATIQCO	Thermo Scientific	SIMPATIQCO	[111]	

Name	Supported instrument vendors	Website/Download	publication	Note
QC-ART		QC-ART	[112]	
Spray Qc	Thermo Scientific, AB SCIEX, extensible to other instrumentation	SprayQc	[113]	
Metriculator		Metriculator	[114]	
Mass QC		MassQC		
Open MS		OpenMS	[115]	

Search result QC

Name	Website/Download/publication	publication	Note
MSStats	MSStats	[116]	can use output from MaxQuant, Proteome Discoverer, Skyline, Progenesis, Spectronaut
MSStatsQC	MSStatsQC	[117]	
PTXQC	PTXQC	[118]	requires MaxQuant search engine output
protti	protti	[119]	

Statistical hypothesis testing

Databases

What are they and where do you get them?

Protein Database Sources and Types

Many mass spectrometry-based proteomic techniques use search algorithms that require a defined theoretical search space to identify peptide sequences based on precursor mass and fragmentation patterns, which are then used to infer the presence and abundance of a protein. The search space is calculated from the potential proteins in a sample, which includes the proteome (often a single species) and expected contaminants. This is called database searching and the flat file of protein sequences in FASTA format acts as a protein database. In this section, we will describe major resources for proteome FASTA files (protein sequence collections), how to retrieve them, and suggested best practices for preserving FASTA file provenance to improve reproducibility.

In general, FASTA sequence collections can be retrieved from three central clearing houses: UniProt, RefSeq, and Ensembl. These will be discussed separately below as they each have specific design goals, data products, and unique characteristics. It is important to learn the following three points for each resource: the source of the underlying data, canonical versus non-canonical sequences, and how versioning works. These points, along with general best practices, such as using a taxonomic identifier, are essential to understand and communicate search settings used in analyses of proteomic datasets. Finally, it is critical to understand that sequence collections from these three resources are not the same, nor do they offer the same sets of species.

Key terminology may vary between resources, so these terms are defined here. The term “taxon identifier” is used across resources and is based on the NCBI taxonomy database. Every taxonomic node has a number, e.g., Homo sapiens (genus species) is 9606 and Mammalia (class) is 40674. This can be useful when retrieving and describing protein sequence collections. Another term used is “annotation”, which has different meanings in different contexts. Broadly, a “genome annotation” is the result of an annotation pipeline to predict coding sequences, and often a gene name/symbol if possible. Two examples are MAKER [**PMID:22192575?**] and the RefSeq annotation pipeline [120]. Alternatively, “protein annotation” (or gene annotation) often refers to the annotation of proteins (gene products) using names and ontology (i.e., protein names, gene names/symbols, functional domains, gene onotology, keywords, etc.). Protein annotation is termed “biocuration” and described in detail by UniProt [121]. Lastly, there are established minimum reporting guidelines for referring to FASTA files established in MIAPE: Mass Spectrometry Informatics that are taxon identifier and number of sequences [122,**PMID:23500130?**]. The FASTA file naming suggestions below are not official but are suggested as a best practice.

UniProt

The Universal Protein Resource (UniProt) [**PMID:14681372?**], has three different products: UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The numerous resources and capabilities associated with the UniProt are not explored in this section, but these are well described on UniProt’s website. UniProtKB is the source of proteomes across the Tree of Life and is the resource we will be describing herein. There are broadly two types of proteome sequence collections: Swiss-Prot/TrEMBL and designated proteomes. The Swiss-Prot/TrEMBL type can be understood by discussing how data is integrated into UniProt. Most protein sequences in UniProt are derived from coding sequences submitted to EMBL-Bank, GenBank and DDBJ. These translated sequences are initially imported into TrEMBL database, which is why TrEMBL is also termed “unreviewed”. There are other sources of protein sequences, as described by UniProt

[123]. These include the Protein Data Bank (PDB), direct protein sequencing, sequences derived from the literature, gene prediction (from sources such as Ensembl) or in-house prediction by UniProt itself. Protein sequences can then be manually curated into the Swiss-Prot database using multiple outlined steps (described in detail by UniProt here [124]) and is why Swiss-Prot is also termed “reviewed”. Note that more than one TrEMBL entry may be removed and replaced by a single Swiss-Prot entry during curation. A search of “organism:9606” at UniProtKB will retrieve both the Swiss-Prot/reviewed and TrEMBL/unreviewed sequences for Homo sapiens. The entries do not overlap, so users often either use just Swiss-Prot or Swiss-Prot combined with TrEMBL, the latter being the most exhaustive option. With ever-increasing numbers of high-quality genome assemblies processed with robust automated annotation pipelines, TrEMBL entries will contain higher quality protein sequences than in the past. In other words, if a mammal species has 20 000 to 40 000 entries in UniProtKB and many of these are TrEMBL, users should be comfortable using all the protein entries to define their search space (more on this later when discussing proteomes at UniProtKB). Determining the expected size of a well-annotated proteome requires additional knowledge, but tools to answer these questions continue to improve. As more and more genome annotations are generated, the backlog of manual curation continues to increase. However, automated genome annotations are also rapidly improving, blurring the line between Swiss-Prot and TrEMBL utility.

The second type of protein sequence collections available at UniProtKB are designated proteomes, with subclasses of “proteome”, “reference proteome” or “pan-proteome”. As defined by UniProt, a proteome is the set of proteins derived from the annotation of a completely sequenced genome assembly (one proteome per genome assembly). This means that a proteome will include both Swiss-Prot and TrEMBL entries present in a single genome annotation, and that all entries in the proteome can be traced to a single complete genome assembly. This aids in tracking provenance as assemblies change, and metrics of these assemblies are available. These metrics include Benchmarking Universal Single-Copy Ortholog (BUSCO) score, and “Completeness” as Standard, Close Standard or Outlier based on the Complete Proteome Detector (CPD). Given the quality of genome annotation pipelines, using a proteome as a FASTA file for a species is the preferred method of defining search spaces now. Outside of humans, no higher eukaryotic Swiss-Prot sequence collections are complete enough for use in proteomics analyses, but this does not mean that the available Swiss-Prot plus TrEMBL protein sequence collection precludes accurate proteomic data analysis. Lastly, the difference between reference proteome and proteome is used to highlight model organisms or organisms of interest, but not to imply improved quality. UniProt also has support for the concept of “pan proteomes” (consensus proteomes for a closely related set of organisms) but this is mostly used for bacteria (e.g., strains of a given species will share a pan proteome).

When retrieving protein sequence collections as Swiss-Prot/TrEMBL or designated proteomes, there is an option of downloading “FASTA (canonical)” or “FASTA (canonical & isoform)”. The later includes additional manually annotated isoforms for Swiss-Prot sequences. Each Swiss-Prot entry has one canonical sequence chosen by the manual curator. Any additional sequence variants (mostly from alternative slicing) are annotated as differences with respect to the canonical sequence. Specifying “canonical” will select only one protein sequence per Swiss-Prot entry while specifying “canonical & isoforms” will download additional protein sequences by including isoforms for Swiss-Prot entries. Recently, an option to “download one protein sequence per gene (FASTA)” has been added. These FASTA files include Swiss-Prot and TrEMBL sequences to number about 20 000 protein sequences for a wide range of higher eukaryotic organisms.

The number of additional isoforms varies considerably by species. In the human, mouse, and rat proteomes of the total number of entries, 26 %, 40 % and 72 % are canonical, respectively. The choice of including isoforms is related to the search algorithm and experimental goals. For instance, if differentiating isoforms is relevant, they should be included otherwise they will not be detected. In cases where isoforms are present in the FASTA (evident by shared protein names) but these cannot be removed prior to downloading (e.g., California sea lion, Zalophus californianus, proteome

UP000515165, release 2022_01), non-redundant FASTA files can be manually generated (i.e., “remove_duplicates.py” via [125]). If possible, retrieving canonical protein sequences via proteomes is the most straight forward approach and in general appropriate for most search algorithms, versus the method of searching and downloading Swiss-Prot and/or TrEMBL entries.

Though FASTA files are the typical input of many search algorithms, UniProt also offers an XML and GFF format download. In contrast to the flat FASTA file format, the XML format includes sequence information as well as associated information like PTMs, which is used in some search algorithms like MetaMorpheus [**PMID:26418581?**].

Once a protein sequence collection has been selected and retrieved, there is the evergreen question of how to name and report this to others in a way that allows them to reproduce the retrieval. The minimum reporting information is the taxon identified and number of sequences used [122,**PMID:23500130?**]. The following naming format (and those below) augments this and is suggested for UniProtKB FASTA files (the use of underscores or hyphens is not critical): [common or scientific name]-[taxon id]-uniprot-[swiss-prot/trembl/proteome]-[UP# if used]-[canonical/canonical plus isoform]-[release] example of a Homo sapiens (human) protein fasta from UniProtKB:

Human-9606-uniprot-proteome-UP000005640-canonical-2022_01.fasta

The importance of the taxon identifier has already been described above and is a consistent identifier across time and shared across resources. The choices of Swiss-Prot and TrEMBL in some combination was discussed above, and Proteome can be “proteome”, “reference proteome” or “pan-proteome”. The proteome identifier (‘UP’ followed by 9 digits) is conserved across releases, and release information should also be included. A confusing issue to newcomers is what the term “release” means. This is a year_month format (e.g., 2022_01), but it is not the date a FASTA file was downloaded or created, nor does it imply there are monthly updates. This release “date” is a traceable release identifier that is listed on UniProt’s website. Including all this information ensures that the exact provenance of a FASTA file is known and allows the FASTA file to be regenerated.

RefSeq

NCBI is a clearing house of numerous types of data and databases. Specific to protein sequence collections, NCBI Reference Sequence Database (RefSeq) provides annotated genomes across the Tree of Life. The newly developed NCBI Datasets portal [126] is the preferred method for accessing the myriad of NCBI data products, though protein sequence collections can also be retrieved from RefSeq directly[127,128]. Like UniProt described above, most of the additional functionality and information available through NCBI Datasets and RefSeq will not be described here, although the Eukaryotic RefSeq annotation dashboard [129] is a noteworthy resource to monitor the progress of new or re-annotations. We recommend exploring the resources available from NCBI [130], utilizing their tutorials and help requests.

RefSeq is akin to the “proteome” sequence collection from UniProtKB, where a release is based on a single genome assembly. If a more complete genome assembly is deposited or additional secondary evidence (e.g., RNA sequencing) is deposited, RefSeq can update the annotation with a new annotation release. Every annotation release will have an annotation report that contains information on the underlying genome assembly, the new genome annotation, secondary evidence used, and various statistics about what was updated. The current annotation release is referred to as the “reference annotation”, but each annotation is numbered sequentially starting at 100 (the first release). Certain species are on scheduled re-annotation, like human and mouse, while other species are updated as needed based on new data and community feedback (ex. release 100 of taxon 9704 was in 2018, but a more contiguous genome assembly resulted in re-annotation to release 101 in

2020). This general process for new and existing species is described in Heck and Neely [PMID:32786681?].

Since RefSeq is genome assembly-centric, its protein sequence collections are retrieved for each species. This contrasts with being able to use a higher-level taxon identifier like 40674 (Mammalia) in UniProt to retrieve a single FASTA. To accomplish this same search in NCBI Datasets requires a Mammalia search, followed by browsing all 2083 genomes and then filtering the results to reference genomes with annotations, and those resulting 188 could be bulk downloaded, though this will still be 188 individual FASTA files. It is possible to download a single FASTA from an upper-level taxon identifier using the NCBI Taxonomy Browser, though this service may be redundant with the new NCBI Datasets portal. Given the constant development of NCBI Datasets, these functionalities may change, but the general RefSeq philosophy of single species FASTA should be kept in mind. Likewise, when retrieving genome annotations there is no ability to specify canonical entries only, but it is possible to use computational tools to remove redundant entries ("remove_duplicates.py" from [125]).

Similar to the UniProtKB FASTA file naming suggestion, the following naming format is suggested for RefSeq protein sequence collection FASTA (the use of underscores or hyphens is not critical): [common or scientific name]-[taxon id]-refseq-[release number] example of a Equus caballus (horse) protein FASTA from RefSeq: Equus_caballus-9796-refseq-103.fasta The release number starts at 100 and is consecutively numbered. Note, the human releases only recently began following this consecutive numbering for Release 110, and previously had a much longer number to be included (e.g., NCBI Release 109.20211119). Also, in a few species (Human and Chinese hamster, currently), there is a reference and an alternate assembly, both with an available annotation. In these cases, including the underlying assembly identifier would be needed. Note that when you retrieve the protein FASTA from NCBI it will include two more identifiers that aren't required in the file name since it can be determined from the taxon identifier and release number. These are the genome assembly used (this is generated by the depositor and follows no naming scheme) and the RefSeq identifier (GCF followed by a number string). These aren't essential for FASTA naming, but are for comparing between UniProt, RefSeq and Ensembl when the same underlying assembly is used (or not, indicating how up to date one is versus the other).

Ensembl

There are two main web portals for Ensembl sequence collections: the Ensembl genome browser [131] has vertebrate organisms and the Ensemble Genome project [132] has specific web portals for different non-vertebrate branches of the Tree of Life. This contrasts with NCBI and UniProt where all branches are centrally available. Recently, Ensembl has created a new portal "Rapid Release" focusing on quickly making annotations available (replacing the "Pre-Ensemble" portal), albeit without the full functionality of the primary Ensembl resources. Overall, Ensembl provides diverse comparative and genomic tools that should be explored, but, specific to this discussion, they provide species-specific genome annotation products similar to RefSeq.

To retrieve a protein sequence collection from Ensemble at any of the portals, a species can be searched using a name, which will then have taxon identifier displayed (but searching by identifier is not readily apparent). From the results you can select your species and follow links for genome annotation. Caution should be used when browsing the annotation products since the protein coding sequence (abbreviated "cds") annotations are nucleic acid sequences (a useable via 3-frame translation if using certain software), while actual translated peptide sequences are in the "pep" folders. The pep folders contain file names with "ab initio" and "all" in the FASTA file names (file extensions are "fa" for FASTA and "gz" indicating gzip compression algorithm), while there may only be one pep product for certain species in the "Rapid Release" portal. The "ab initio" FASTA files contain mostly predicted gene products. The "all" FASTA files are the usable protein sequence collections. Ensembl FASTA files usually have some protein sequence redundancy.

Ensembl provides a release number for all the databases within each portal. Similar to the UniProt file naming suggestion, the following naming format is suggested for Ensembl protein sequence collection FASTA (the use of underscores or hyphens is not critical):

[common or scientific name]-[taxon id]-ensembl-[abinitio/all]-[rapid]-[release number]

example of a Sus scrofa (pig) protein FASTA from Ensembl:

Pig-9823-ensembl-all-106.fasta

Similar to the FASTA download from RefSeq, the downloaded file name can include additional identifying information related to the underlying genome assembly. Again, this is not required for labeling, but is useful to easily compare assembly versions.

Since much of the data from Ensembl is also regularly processed into UniProt, using UniProt sequence collections instead may be preferred. That said, they are not on the same release schedule nor will the FASTA files contain the same proteins. Ensembl sequences still must go through the established protein sequence pipeline at UniProt to remove redundancy and conform to UniProt accession and FASTA header formats. Moreover, the gene-centric and comparative tools built into Ensembl may be more experimentally appropriate and using an Ensembl protein sequence collection can better leverage those tools.

Other resources

There are other locations of protein sequence collections, and these will likewise have different FASTA file formatting; sequences may have unusual characters, and formats of accessions and FASTA header lines may need to be reformatted to be compatible with search software. These alternatives include institutes like the Joint Genome Institute's microbial genome clearing house, species-specific community resource (e.g., PomBase, FlyBase, WormBase, TryTrypDB, etc.), and one-off websites tenuously hosting in-house annotations. It is preferred to use protein sequence collection from the main three sources described here, since provenance can be tracked, and versions maintained. It is beyond the scope of this discussion to address other genome annotation resources, how they are versioned, or the best way to describe FASTA files retrieved from those sources. In these cases, defaulting to the minimum requirements of listing number of entries and supplying the FASTA along with data are necessary.

Contaminants

Samples are rarely comprised of only proteins from the species of interest. There can be protein contamination during sample collection or processing. This may include proteins from human skin, wool from clothing, particles from latex, or even porcine trypsin itself, all of which contain proteins that can be digested along with the intended sample and analyzed in the mass spectrometer. Avoiding unwanted matching of mass spectra originating from contaminant proteins to the cellular proteins due to sequence similarities is important to the identification and quantitation of as many cellular proteins as possible. To avoid random matching, repositories of supplementary sequences for contaminant proteins have been added to a reference database for MS data searches. Appending a contaminants database to the reference database allows the identification of peptides that are not exclusive to one species. Peptides that are exclusive to the organism of interest are used to calculate abundance to avoid inflated quantitative results due to potential contaminant peptides.

As early as 2004, The Global Proteome Machine was providing a protein sequence collection of these common Repository of Adventitious Proteins (cRAP), while another contaminant list was published in

2008 [PMID:18790129?]. The current cRAP version (v1.0) was described in 2012 [133] and is still widely in use today. cRAP is the contaminant protein list used in nearly all modern database searching software, though the documentation, versioning or updating of many of these “built-in” contaminant sequence collections is difficult to follow. There is also another contaminant sequence collection distributed with MaxQuant. Together, the cRAP and MaxQuant contaminant protein sequence collections are found in some form across most software, including MetaMorpheus and Philosopher (available in FragPipe) [PMID:32669682?]. This list of known frequently contaminating proteins can either be automatically included by the software or can be retrieved as a FASTA to be used along with the primary search FASTA(s). Recently the Hao Lab has revisited these common contaminant sequences in an effort to update the protein sequences, test their utility on experimental data, and add or remove entries [134].

In addition to these environmentally unintended contaminants, there are known contaminants that also have available protein sequence collections (or can be generated using the steps above) and should be included in the search space. These can include the media cells were grown in (e.g., fetal bovine serum [PMID:20641139?,PMID:33532042?], food fed to cells/animals (e.g., Caenorhabditis elegans grown on Escherichia coli) or known non-specific binders in affinity purification (i.e., CRAPome [PMID:23921808?]). The common Repository of Fetal Bovine Serum Proteins (cRFP)[PMID:31475827?] are protein lists of common protein contaminants and fetal serum bovine sequences used to reduced the number of falsely identified proteins in cell culture experiments. Cells washed or cultured in contaminant free media before harvest or the collection of secreted proteins depletes most high abundance contaminant proteins but the sequence similarity between contaminant and secreted proteins can cause false identifications and overestimation of the true protein abundance leading to wasted resources and time on validating false leads. As emphasized throughout this section, accurately defining the search space is essential for accurate results and, especially in the case of contaminants, requires knowledge of the experiment and sample processing to adequately define possible background proteins.

Choosing the right database

Proteomics data analysis requires carefully matching the search space (defined by the database choice) with the expected proteins. A properly chosen database will minimize false positives and false negatives. Choosing a database that is too large will increase the number of false positives, or decoy hits, which in turn will reduce the total number of identifiable proteins. For this reason it is ill advised to search against all possible protein sequences ever predicted from any genomic sequence. On the other hand, choosing a database that is too small may increase false negatives, or missed protein identifications, because in order for a protein to be identified it must be present in the database. Thus, proteomics practitioners must do their best to predict the proteins that might be in their sample before they analyze their data.

Proteomics data analysis requires carefully aligning the search space with the expected proteome and the statistical approach of the search algorithm. Search algorithms can self-correct when a database is overly large such that higher identity thresholds are required for identification to minimize false positives (e.g., Mascot), while smaller experiment-specific search spaces (also referred to as “subsets”) can have unintended effects on false positives if not managed appropriately [PMID:34236864?,PMID:30560673?,PMID:26125591?] or may even improve protein identifications [PMID:27975281?]. Whether to employ a search space that is sample-specific (i.e., subset), species-specific (with only canonical proteins, described below), exhaustive species-specific (including all isoforms), or even larger clade-level protein sequence set (e.g., the over 14 million protein sequences associated with Fungi, taxon identifier 4751) is a complex issue that is experiment and software dependent. Moreover, in cases where no species-specific protein sequence collection exists, homology-based searching can be used (as described in [PMID:32786681?]). In each of these cases, proteomics practitioners must understand their specific experimental sample and search algorithm in

order to know how to best define the search space, which is essential to yielding accurate results. See more discussion of database choice in the following section.

Biological Interpretation

1. term enrichment analysis (KEGG, GO)
2. network analysis methods
3. structure analysis
4. isoform analysis
5. follow-up experiments

Methods for protein or peptide fractionation

Protein fractionation * SDS-PAGE

Peptide fractionation * bRP

Tandem Mass Spectrometry and Peptide Fragmentation

Tandem Mass Spectrometry

Why do we need tandem mass spectra data for proteomics? As mass accuracy and resolution increase, the number of potential peptide matches decrease. Still, there are many potential matches to a human peptide with a 5 ppm mass tolerance window (citation). Additional information is required to uniquely assign a signal to a peptide. Early approaches described use of accurate mass and time data to uniquely identify peptides [135]. Most current research, however, has pursued using the pattern of peptide fragment masses helps narrow down the number of potential peptides.

Peptide Fragmentation

how it works Modern proteomics utilizes hybrid mass spectrometers, where hybrid means there are multiple mass analyzers working in harmony. Most hybrid mass spectrometers combine a quadrople for mass selection before a high resolution analyzer, usually a TOF or an orbitrap. In these hybrid mass spectrometers, mass selection happens using the first quadrupole to isolate a small mass range around that mass (usually less than 1 thompson). The selected ions then pass to the collision cell where fragmentation happens.

1. Nomenclature of backbone fragments (abc, xyz, internal, losses).
2. Mechanisms of peptide fragmentation, collision based (CID and HCD, PQD?) versus electron based (ECD and ETD).

Basic Spectral Interpretation

Segway into the section on raw data analysis.

Experiment Design

This section should discuss trade offs and balancing them to design an experiment. 1. constraints: Each experiment will have different constraints, which may include the number of samples needed for analysis, or desire to quantify a specific subset of proteins within a sample. 2. sample size 3. statistics 4. costs

Acknowledgements

The authors thank Phil Wilmarth for helpful input. Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

References

1. **Electrospray Ionization for Mass Spectrometry of Large Biomolecules**
John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, Craig M Whitehouse
Science (1989-10-06) https://doi.org/cq2q43
DOI: 10.1126/science.2675315 · PMID: 2675315

2. **Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry**
Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, T Matsuo
Rapid Communications in Mass Spectrometry (1988-08) https://doi.org/ffbwwr
DOI: 10.1002/rcm.1290020802

3. **An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics**
Dirk A Wolters, Michael P Washburn, John R Yates
Analytical Chemistry (2001-10-25) https://doi.org/bn4kq6
DOI: 10.1021/ac010617e · PMID: 11774908

4. **A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry**
Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, Ruedi Aebersold
Analytical Chemistry (2003-07-15) https://doi.org/b2xv45
DOI: 10.1021/ac0341261 · PMID: 14632076

5. **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**
Joshua E Elias, Steven P Gygi
Nature Methods (2007-03) https://doi.org/djz7fz
DOI: https://doi.org/10.1038/nmeth1019

6. **Mass-spectrometric exploration of proteome structure and function**
Ruedi Aebersold, Matthias Mann
Nature (2016-09) https://doi.org/f83zqm
DOI: 10.1038/nature19949 · PMID: 27629641

7. **High-throughput quantitative top-down proteomics**
Kellye A Cupp-Sutton, Si Wu
Molecular Omics (2020) https://doi.org/gnx98p
DOI: 10.1039/c9mo00154a · PMID: 31932818 · PMCID: PMC7529119

8. **Proteoforms as the next proteomics currency**
Lloyd M Smith, Neil L Kelleher
Science (2018-03-09) https://doi.org/gn6p4x
DOI: 10.1126/science.aat1884 · PMID: 29590032 · PMCID: PMC5944612

9. **The quantitative proteome of a human cell line**
Martin Beck, Alexander Schmidt, Johan Malmstroem, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, Ruedi Aebersold
Molecular Systems Biology (2011-01) https://doi.org/drjmfj
DOI: 10.1038/msb.2011.82 · PMID: 22068332 · PMCID: PMC3261713

10. **A Quantitative Proteome Map of the Human Body**

Lihua Jiang, Meng Wang, Shin Lin, Ruiqi Jian, Xiao Li, Joanne Chan, Guanlan Dong, Huaying Fang, Aaron E Robinson, Michael P Snyder, … Simona Volpi
Cell (2020-10) https://doi.org/ghbjrk
DOI: 10.1016/j.cell.2020.08.036 · PMID: 32916130 · PMCID: PMC7575058

11. **Proteoform: a single term describing protein complexity**
Lloyd M Smith, Neil L Kelleher
Nature Methods (2013-02-27) https://doi.org/gwcc
DOI: 10.1038/nmeth.2369 · PMID: 23443629 · PMCID: PMC4114032

12. **Post-translational modifications in proteins: resources, tools and prediction methods**
Shahin Ramazi, Javad Zahiri
Database (2021-01-01) https://doi.org/gpjs7g
DOI: 10.1093/database/baab012 · PMID: 33826699 · PMCID: PMC8040245

13. **Deciphering protein post-translational modifications using chemical biology tools**
Anne C Conibear
Nature Reviews Chemistry (2020-10-06) https://doi.org/gqnxhb
DOI: 10.1038/s41570-020-00223-8

14. **Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions**
Christopher J Oldfield, AKeith Dunker
Annual Review of Biochemistry (2014-06-02) https://doi.org/gfw6t4
DOI: 10.1146/annurev-biochem-072711-164947 · PMID: 24606139

15. **A Review: Molecular Chaperone-mediated Folding, Unfolding and Disaggregation of Expressed Recombinant Proteins**
Komal Fatima, Fatima Naqvi, Hooria Younas
Cell Biochemistry and Biophysics (2021-02-25) https://doi.org/gpd4pb
DOI: 10.1007/s12013-021-00970-5 · PMID: 33634426

16. **Guide to protein purification**
Richard R Burgess, Murray P Deutscher
Elsevier/Academic Press (2009)
ISBN: 9780123745361

17. **The protein protocols handbook**
John M Walker (editor)
Humana Press (2009)
ISBN: 9781597451987

18. **An overview of cell disruption methods for intracellular biomolecules recovery**
Tatiane Aparecida Gomes, Cristina Maria Zanette, Michele Rigon Spier
Preparative Biochemistry & Biotechnology (2020-02-19) https://doi.org/gqpdr5
DOI: 10.1080/10826068.2020.1728696 · PMID: 32074000

19. **Fast and Sensitive Total Protein and Peptide Assays for Proteomic Analysis**
Jacek R Wiśniewski, Fabienne Z Gaugaz
Analytical Chemistry (2015-04-09) https://doi.org/f3nsk2
DOI: 10.1021/ac504689z · PMID: 25837572

20. **Getting intimate with trypsin, the leading protease in proteomics**
Elien Vandermarliere, Michael Mueller, Lennart Martens
Mass Spectrometry Reviews (2013-06-15) https://doi.org/gn64qb
DOI: 10.1002/mas.21376 · PMID: 23775586

21. **In SilicoProteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage**
Jesse G Meyer
ISRN Computational Biology (2014-04-22) https://doi.org/gb6s2r
DOI: 10.1155/2014/960902 · PMID: 30687733 · PMCID: PMC6347401

22. **Venomics and antivenomics of Indian spectacled cobra (Naja naja) from the Western Ghats**
Muralidharan Vanuopadath, Dileepkumar Raveendran, Bipin Gopalakrishnan Nair, Sudarslal Sadasivan Nair
Acta Tropica (2022-04) https://doi.org/gpbzf7
DOI: 10.1016/j.actatropica.2022.106324 · PMID: 35093326

23. **Sequencing-Grade De novo Analysis of MS/MS Triplets (CID/HCD/ETD) From Overlapping Peptides**
Adrian Guthals, Karl R Clauser, Ari M Frank, Nuno Bandeira
Journal of Proteome Research (2013-05-30) https://doi.org/f47kqd
DOI: 10.1021/pr400173d · PMID: 23679345 · PMCID: PMC4591044

24. **Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data**
Rachel M Miller, Robert J Millikin, Connor V Hoffmann, Stefan K Solntsev, Gloria M Sheynkman, Michael R Shortreed, Lloyd M Smith
Journal of Proteome Research (2019-08-04) https://doi.org/gpqmp2
DOI: 10.1021/acs.jproteome.9b00330 · PMID: 31378069 · PMCID: PMC6733628

25. **Expanding Proteome Coverage with Orthogonal-specificity α-Lytic Proteases**
Jesse G Meyer, Sangtae Kim, David A Maltby, Majid Ghassemian, Nuno Bandeira, Elizabeth A Komives
Molecular & Cellular Proteomics (2014-03) https://doi.org/f5vgcg
DOI: 10.1074/mcp.m113.034710 · PMID: 24425750 · PMCID: PMC3945911

26. **Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides**
Philipp T Kaulich, Liam Cassidy, Jürgen Bartel, Ruth A Schmitz, Andreas Tholey
Journal of Proteome Research (2021-03-24) https://doi.org/gpqmpz
DOI: 10.1021/acs.jproteome.1c00115 · PMID: 33760615

27. **A Multiple Protease Strategy to Optimise the Shotgun Proteomics of Mature Medicinal Cannabis Buds**
Delphine Vincent, Vilnis Ezernieks, Simone Rochfort, German Spangenberg
International Journal of Molecular Sciences (2019-11-11) https://doi.org/gpqmp3
DOI: 10.3390/ijms20225630 · PMID: 31717952 · PMCID: PMC6888629

28. **Confetti: A Multiprotease Map of the HeLa Proteome for Comprehensive Proteomics**
Xiaofeng Guo, David C Trudgian, Andrew Lemoff, Sivaramakrishna Yadavalli, Hamid Mirzaei
Molecular & Cellular Proteomics (2014-06) https://doi.org/f56bwx
DOI: 10.1074/mcp.m113.035170 · PMID: 24696503 · PMCID: PMC4047476

29. **An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas**
Piero Giansanti, Thin Thin Aye, Henk van den Toorn, Mao Peng, Bas van Breukelen, Albert JR Heck
Cell Reports (2015-06) https://doi.org/gpqmpx
DOI: 10.1016/j.celrep.2015.05.029 · PMID: 26074081

30. **The Nobel Prize in Chemistry 1946**

NobelPrize.org
https://www.nobelprize.org/prizes/chemistry/1946/speedread/

31. **Online, High-Pressure Digestion System for Protein Characterization by Hydrogen/Deuterium Exchange and Mass Spectrometry**
Lisa M Jones, Hao Zhang, Ilan Vidavsky, Michael L Gross
Analytical Chemistry (2010-01-22) https://doi.org/b993rm
DOI: 10.1021/ac902477u · PMID: 20095571 · PMCID: PMC2826105

32. **Hydrogen/deuterium exchange in mass spectrometry**
Yury Kostyukevich, Thamina Acter, Alexander Zherebker, Arif Ahmed, Sunghwan Kim, Eugene Nikolaev
Mass Spectrometry Reviews (2018-03-30) https://doi.org/gffzx8
DOI: 10.1002/mas.21565 · PMID: 29603316

33. **Proteinase K**
W Saenger
Handbook of Proteolytic Enzymes (2013) https://doi.org/gkfkcz
DOI: https://doi.org/10.1016/b978-0-12-382219-2.00714-6 · ISBN: 9780123822192

34. **Insights into protein post-translational modification landscapes of individual human cells by trapped ion mobility time-of-flight mass spectrometry**
Benjamin C Orsburn, Yuting Yuan, Namandjé N Bumpus
Nature Communications (2022-11-25) https://doi.org/grhnzj
DOI: 10.1038/s41467-022-34919-w · PMID: 36433961 · PMCID: PMC9700839

35. **TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing**
Jiaming Li, Zhenying Cai, Ryan D Bomgarden, Ian Pike, Karsten Kuhn, John C Rogers, Thomas M Roberts, Steven P Gygi, Joao A Paulo
Journal of Proteome Research (2021-04-26) https://doi.org/gjt6rs
DOI: 10.1021/acs.jproteome.1c00168 · PMID: 33900084 · PMCID: PMC8210943

36. **Simultaneous Quantification of the Acetylome and Succinylome by 'One-Pot' Affinity Enrichment**
Nathan Basisty, Jesse G Meyer, Lei Wei, Bradford W Gibson, Birgit Schilling
PROTEOMICS (2018-08-19) https://doi.org/gn4cmb
DOI: 10.1002/pmic.201800123 · PMID: 30035354 · PMCID: PMC6175148

37. **Bacterial cellulose nanofibers for albumin depletion from human serum**
Emel Tamahkar, Ceyhun Babaç, Tülin Kutsal, Erhan Pişkin, Adil Denizli
Process Biochemistry (2010-10) https://doi.org/bzqhtf
DOI: 10.1016/j.procbio.2010.07.007

38. **Electrostatic Repulsion Hydrophilic Interaction Chromatography for Isocratic Separation of Charged Solutes and Selective Isolation of Phosphopeptides**
Andrew J Alpert
Analytical Chemistry (2007-11-21) https://doi.org/bt84c3
DOI: 10.1021/ac070997p · PMID: 18027909

39. **Novel Application of Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) in Shotgun Proteomics: Comprehensive Profiling of Rat Kidney Proteome**
Piliang Hao, Tiannan Guo, Xin Li, Sunil S Adav, Jie Yang, Meng Wei, Siu Kwan Sze
Journal of Proteome Research (2010-06-10) https://doi.org/bgkxkx
DOI: 10.1021/pr100037h · PMID: 20450224

40. **The Nobel Prize in Chemistry 2002**
NobelPrize.org
https://www.nobelprize.org/prizes/chemistry/2002/summary/

41. **Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules**
Michael Karas, Doris Bachmann, Franz Hillenkamp
Analytical Chemistry (1985-12-01) https://pubs.acs.org/doi/abs/10.1021/ac00291a042
DOI: 10.1021/ac00291a042

42. **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**
Michael Karas, Franz Hillenkamp
Analytical Chemistry (1988-10-15) https://doi.org/d577jp
DOI: 10.1021/ac00171a028 · PMID: 3239801

43. **The Scientist :: Nobel Prize controversy** (2007-05-17)
https://web.archive.org/web/20070517202246/http://cmbi.bjmu.edu.cn/news/0212/55.htm

44. **α-Cyano-4-hydroxycinnamic acid, sinapinic acid, and ferulic acid as matrices and alkylating agents for matrix-assisted laser desorption/ionization time-of-flight mass spectrometric analysis of cysteine-containing peptides**
Hongmei Yang, Debin Wan, Fengrui Song, Zhiqiang Liu, Shuying Liu
Rapid communications in mass spectrometry: RCM (2013-06-30)
https://pubmed.ncbi.nlm.nih.gov/23681820
DOI: 10.1002/rcm.6587

45. **The Desorption Process in MALDI**
Klaus Dreisewerd
Chemical Reviews (2003-01-24) https://doi.org/cpzqmq
DOI: 10.1021/cr010375i · PMID: 12580636

46. **Matrix Dependence of Metastable Fragmentation of Glycoproteins in MALDI TOF Mass Spectrometry**
Michael Karas, Ute Bahr, Kerstin Strupat, Franz Hillenkamp, Anthony Tsarbopoulos, Birendra N Pramanik
Analytical Chemistry (1995-02-01) https://doi.org/b54gnt
DOI: 10.1021/ac00099a029

47. **XX. On the equilibrium of liquid conducting masses charged with electricity**
Lord Rayleigh
The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science (1882-09)
https://doi.org/c6bp6h
DOI: 10.1080/14786448208628425

48. **Capillary Liquid Chromatography/Mass Spectrometry for Peptide and Protein Characterization**
MT Davis, DC Stahl, KM Swiderek, TD Lee
Methods (1994-09) https://doi.org/fmtw5k
DOI: 10.1006/meth.1994.1031

49. **Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database**
John R Yates, Jimmy K Eng, Ashley L McCormack, David Schieltz
Analytical Chemistry (1995-04-15) https://doi.org/dtcrm9
DOI: 10.1021/ac00104a020 · PMID: 7741214

50. **Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ**
Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, Matthias Mann
Molecular & Cellular Proteomics (2014-09) https://doi.org/f6hmm5
DOI: 10.1074/mcp.m113.031591 · PMID: 24942700 · PMCID: PMC4159666

51. **The MaxQuant computational platform for mass spectrometry-based shotgun proteomics**
Stefka Tyanova, Tikira Temu, Juergen Cox
Nature Protocols (2016-10-27) https://doi.org/f89fwn
DOI: 10.1038/nprot.2016.136 · PMID: 27809316

52. **Advances in proteomics data analysis and display using an accurate mass and time tag approach**
Jennifer SD Zimmer, Matthew E Monroe, Wei-Jun Qian, Richard D Smith
Mass Spectrometry Reviews (2006) https://doi.org/bqv76w
DOI: 10.1002/mas.20071 · PMID: 16429408 · PMCID: PMC1829209

53. **A New Algorithm Using Cross-Assignment for Label-Free Quantitation with LC−LTQ-FT MS**
Victor P Andreev, Lingyun Li, Lei Cao, Ye Gu, Tomas Rejtar, Shiaw-Lin Wu, Barry L Karger
Journal of Proteome Research (2007-04-19) https://doi.org/d3fshp
DOI: 10.1021/pr0606880 · PMID: 17441747 · PMCID: PMC2563808

54. **IDEAL-Q, an Automated Tool for Label-free Quantitation Analysis Using an Efficient Peptide Alignment Approach and Spectral Data Validation**
Chih-Chiang Tsou, Chia-Feng Tsai, Ying-Hao Tsui, Putty-Reddy Sudhir, Yi-Ting Wang, Yu-Ju Chen, Jeou-Yuan Chen, Ting-Yi Sung, Wen-Lian Hsu
Molecular & Cellular Proteomics (2010-01) https://doi.org/d9gxpm
DOI: 10.1074/mcp.m900177-mcp200 · PMID: 19752006 · PMCID: PMC2808259

55. ** SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling**
Lukas N Mueller, Oliver Rinner, Alexander Schmidt, Simon Letarte, Bernd Bodenmiller, Mi-Youn Brusniak, Olga Vitek, Ruedi Aebersold, Markus Müller
PROTEOMICS (2007-08-28) https://doi.org/c5zp29
DOI: 10.1002/pmic.200700057 · PMID: 17726677

56. **Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model**
Matthew Y Lim, João A Paulo, Steven P Gygi
Journal of Proteome Research (2019-09-23) https://doi.org/ggd6nh
DOI: 10.1021/acs.jproteome.9b00492 · PMID: 31547658 · PMCID: PMC7346880

57. **IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs**
Fengchao Yu, Sarah E Haynes, Alexey I Nesvizhskii
Molecular & Cellular Proteomics (2021) https://doi.org/gqfrh3
DOI: 10.1016/j.mcpro.2021.100077 · PMID: 33813065 · PMCID: PMC8131922

58. **Post Analysis Data Acquisition for the Iterative MS/MS Sampling of Proteomics Mixtures**
Michael R Hoopmann, Gennifer E Merrihew, Priska D von Haller, Michael J MacCoss
Journal of Proteome Research (2009-03-03) https://doi.org/fgtwdr
DOI: 10.1021/pr800828p · PMID: 19256536 · PMCID: PMC2671646

59. **The Implications of Proteolytic Background for Shotgun Proteomics**

Paola Picotti, Ruedi Aebersold, Bruno Domon
Molecular & Cellular Proteomics (2007-09) https://doi.org/ffrsx7
DOI: 10.1074/mcp.m700029-mcp200 · PMID: 17533221

60. **Large-scale characterization of HeLa cell nuclear phosphoproteins**
Sean A Beausoleil, Mark Jedrychowski, Daniel Schwartz, Joshua E Elias, Judit Villén, Jiaxu Li, Martin A Cohn, Lewis C Cantley, Steven P Gygi
Proceedings of the National Academy of Sciences (2004-08-09) https://doi.org/cjmg8h
DOI: 10.1073/pnas.0404720101 · PMID: 15302935 · PMCID: PMC514446

61. **Evaluation of the utility of neutral-loss-dependent MS3 strategies in large-scale phosphorylation analysis**
Judit Villén, Sean A Beausoleil, Steven P Gygi
PROTEOMICS (2008-11) https://doi.org/c52q48
DOI: 10.1002/pmic.200800283 · PMID: 18972524 · PMCID: PMC2745099

62. **Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of N-Linked Glycoproteins**
Charandeep Singh, Cleidiane G Zampronio, Andrew J Creese, Helen J Cooper
Journal of Proteome Research (2012-08-06) https://doi.org/f36tfq
DOI: 10.1021/pr300257c · PMID: 22800195

63. **Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques**
Jesse G Meyer, Birgit Schilling
Expert Review of Proteomics (2017-05-04) https://doi.org/gk6gdn
DOI: 10.1080/14789450.2017.1322904 · PMID: 28436239 · PMCID: PMC5671767

64. **Review of software tools for design and analysis of large scale MRM proteomic datasets**
Christopher M Colangelo, Lisa Chung, Can Bruce, Kei-Hoi Cheung
Methods (2013-06) https://doi.org/f449z8
DOI: 10.1016/j.ymeth.2013.05.004 · PMID: 23702368 · PMCID: PMC3775261

65. **Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and13C-labeled peptide standards**
Eric Kuhn, Jiang Wu, Johann Karl, Hua Liao, Werner Zolg, Brad Guild
PROTEOMICS (2004-04) https://doi.org/cxn7fw
DOI: 10.1002/pmic.200300670 · PMID: 15048997

66. **Mass Spectrometric Quantitation of Peptides and Proteins Using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA)**
NLeigh Anderson, Norman G Anderson, Lee R Haines, Darryl B Hardie, Robert W Olafson, Terry W Pearson
Journal of Proteome Research (2004-02-06) https://doi.org/c494px
DOI: 10.1021/pr034086h · PMID: 15113099

67. **Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins**
Leigh Anderson, Christie L Hunter
Molecular & Cellular Proteomics (2006-04) https://doi.org/c8f5g7
DOI: 10.1074/mcp.m500331-mcp200 · PMID: 16332733

68. **A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition**

Veronika Vidova, Zdenek Spacil
Analytica Chimica Acta (2017-04) https://doi.org/f93hpx
DOI: 10.1016/j.aca.2017.01.059 · PMID: 28351641

69. **The current status of clinical proteomics and the use of MRM and MRM³for biomarker validation**
Jérôme Lemoine, Tanguy Fortin, Arnaud Salvador, Aurore Jaffuel, Jean-Philippe Charrier, Geneviève Choquet-Kastylevsky
Expert Review of Molecular Diagnostics (2012-05) https://doi.org/f337q3
DOI: 10.1586/erm.12.32 · PMID: 22616699

70. **High Sensitivity Detection of Plasma Proteins by Multiple Reaction Monitoring of N-Glycosites**
Jianru Stahl-Zeng, Vinzenz Lange, Reto Ossola, Katrin Eckhardt, Wilhelm Krek, Ruedi Aebersold, Bruno Domon
Molecular & Cellular Proteomics (2007-10) https://doi.org/bwgf39
DOI: 10.1074/mcp.m700132-mcp200 · PMID: 17644760

71. **A database of mass spectrometric assays for the yeast proteome**
Paola Picotti, Henry Lam, David Campbell, Eric W Deutsch, Hamid Mirzaei, Jeff Ranish, Bruno Domon, Ruedi Aebersold
Nature Methods (2008-11) https://doi.org/dvs3c7
DOI: 10.1038/nmeth1108-913 · PMID: 18974732 · PMCID: PMC2770732

72. **Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics**
Amelia C Peterson, Jason D Russell, Derek J Bailey, Michael S Westphall, Joshua J Coon
Molecular & Cellular Proteomics (2012-11) https://doi.org/f4gzw2
DOI: 10.1074/mcp.o112.020131 · PMID: 22865924 · PMCID: PMC3494192

73. **Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics**
Graziella E Ronsein, Nathalie Pamir, Priska D von Haller, Daniel S Kim, Michael N Oda, Gail P Jarvik, Tomas Vaisar, Jay W Heinecke
Journal of Proteomics (2015-01) https://doi.org/f6wq8n
DOI: 10.1016/j.jprot.2014.10.017 · PMID: 25449833 · PMCID: PMC4259393

74. **Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer**
Samuel Purvine, Jason-Thomas Eppel*, Eugene C Yi, David R Goodlett
PROTEOMICS (2003-06) https://doi.org/c37bqm
DOI: 10.1002/pmic.200300362 · PMID: 12833507

75. **Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra**
John D Venable, Meng-Qiu Dong, James Wohlschlegel, Andrew Dillin, John R Yates III
Nature Methods (2004-09-29) https://doi.org/dm8rm4
DOI: 10.1038/nmeth705 · PMID: 15782151

76. **UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation**
Robert S Plumb, Kelly A Johnson, Paul Rainville, Brian W Smith, Ian D Wilson, Jose M Castro-Perez, Jeremy K Nicholson
Rapid Communications in Mass Spectrometry (2006) https://doi.org/c8gx39
DOI: 10.1002/rcm.2550 · PMID: 16755610

77. **Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean**
Alexandre Panchaud, Alexander Scherl, Scott A Shaffer, Priska D von Haller, Hemantha D Kulasekara, Samuel I Miller, David R Goodlett
Analytical Chemistry (2009-07-02) https://doi.org/dwbrjn
DOI: 10.1021/ac900888s · PMID: 19572557 · PMCID: PMC3086478

78. **Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation**
Tamar Geiger, Juergen Cox, Matthias Mann
Molecular & Cellular Proteomics (2010-10) https://doi.org/fqrb3c
DOI: 10.1074/mcp.m110.001537 · PMID: 20610777 · PMCID: PMC2953918

79. **XDIA: improving on the label-free data-independent analysis**
Paulo C Carvalho, Xuemei Han, Tao Xu, Daniel Cociorva, Maria da Gloria Carvalho, Valmir C Barbosa, John R Yates III
Bioinformatics (2010-01-26) https://doi.org/cn8x3r
DOI: 10.1093/bioinformatics/btq031 · PMID: 20106817 · PMCID: PMC2832823

80. **ETISEQ – an algorithm for automated elution time ion sequencing of concurrently fragmented peptides for mass spectrometry-based proteomics**
Jason WH Wong, Alexander B Schwahn, Kevin M Downard
BMC Bioinformatics (2009-08-10) https://doi.org/c3rshm
DOI: 10.1186/1471-2105-10-244 · PMID: 19664259 · PMCID: PMC2731054

81. **Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis**
Ludovic C Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, Ruedi Aebersold
Molecular & Cellular Proteomics (2012-06) https://doi.org/fzqdc5
DOI: 10.1074/mcp.o111.016717 · PMID: 22261725 · PMCID: PMC3433915

82. **Multiplexed MS/MS for improved data-independent acquisition**
Jarrett D Egertson, Andreas Kuehn, Gennifer E Merrihew, Nicholas W Bateman, Brendan X MacLean, Ying S Ting, Jesse D Canterbury, Donald M Marsh, Markus Kellmann, Vlad Zabrouskov, … Michael J MacCoss
Nature Methods (2013-06-23) https://doi.org/gjqdct
DOI: 10.1038/nmeth.2528 · PMID: 23793237 · PMCID: PMC3881977

83. **Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results**
Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Yue Xuan, Julia Sondermann, Manuela Schmidt, David Gomez-Varela, Lukas Reiter
Molecular & Cellular Proteomics (2017-12) https://doi.org/gcqvm4
DOI: 10.1074/mcp.ra117.000314 · PMID: 29070702 · PMCID: PMC5724188

84. **diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition**
Florian Meier, Andreas-David Brunner, Max Frank, Annie Ha, Isabell Bludau, Eugenia Voytik, Stephanie Kaspar-Schoenefeld, Markus Lubeck, Oliver Raether, Nicolai Bache, … Matthias Mann
Nature Methods (2020-11-30) https://doi.org/gmhqdm
DOI: 10.1038/s41592-020-00998-0 · PMID: 33257825

85. **Single-Shot 10K Proteome Approach: Over 10,000 Protein Identifications by Data-Independent Acquisition-Based Single-Shot Proteomics with Ion Mobility Spectrometry**

Yusuke Kawashima, Hirotaka Nagai, Ryo Konno, Masaki Ishikawa, Daisuke Nakajima, Hironori Sato, Ren Nakamura, Tomoyuki Furuyashiki, Osamu Ohara
Journal of Proteome Research (2022-05-06) https://doi.org/gqgg5x
DOI: 10.1021/acs.jproteome.2c00023 · PMID: 35522919 · PMCID: PMC9171847

86. **In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics**
Enrique Audain, Julian Uszkoreit, Timo Sachsenberg, Julianus Pfeuffer, Xiao Liang, Henning Hermjakob, Aniel Sanchez, Martin Eisenacher, Knut Reinert, David L Tabb, … Yasset Perez-Riverol
Journal of Proteomics (2017-01) https://doi.org/f9r8r6
DOI: 10.1016/j.jprot.2016.08.002 · PMID: 27498275

87. **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**
Jürgen Cox, Matthias Mann
Nature Biotechnology (2008-11-30) https://doi.org/crn24x
DOI: 10.1038/nbt.1511 · PMID: 19029910

88. **MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics**
Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, Alexey I Nesvizhskii
Nature Methods (2017-04-10) https://doi.org/f9z6p7
DOI: 10.1038/nmeth.4256 · PMID: 28394336 · PMCID: PMC5409104

89. **MS-GF+ makes progress towards a universal database search tool for proteomics**
Sangtae Kim, Pavel A Pevzner
Nature Communications (2014-10-31) https://doi.org/ggkdq8
DOI: 10.1038/ncomms6277 · PMID: 25358478 · PMCID: PMC5036525

90. **A method for reducing the time required to match protein sequences with tandem mass spectra**
Robertson Craig, Ronald C Beavis
Rapid Communications in Mass Spectrometry (2003) https://doi.org/b7bgb9
DOI: 10.1002/rcm.1198 · PMID: 14558131

91. **TANDEM: matching proteins with tandem mass spectra**
R Craig, RC Beavis
Bioinformatics (2004-02-19) https://doi.org/cthw6n
DOI: 10.1093/bioinformatics/bth092 · PMID: 14976030

92. **Skyline: an open source document editor for creating and analyzing targeted proteomics experiments**
Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, Michael J MacCoss
Bioinformatics (2010-02-09) https://doi.org/bqx9rq
DOI: 10.1093/bioinformatics/btq054 · PMID: 20147306 · PMCID: PMC2844992

93. **DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput**
Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, Markus Ralser
Nature Methods (2019-11-25) https://doi.org/gj9xgj
DOI: 10.1038/s41592-019-0638-x · PMID: 31768060 · PMCID: PMC6949130

94. **Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry**
Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, Ming Li
Nature Methods (2018-12-20) https://doi.org/gftvmn
DOI: 10.1038/s41592-018-0260-3 · PMID: 30573815

95. **Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues**
Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Saša M Miladinović, Lin-Yang Cheng, Simon Messner, Tobias Ehrenberger, Vito Zanotelli, Yulia Butscheid, Claudia Escher, … Lukas Reiter
Molecular & Cellular Proteomics (2015-05) https://doi.org/f7b76h
DOI: 10.1074/mcp.m114.044305 · PMID: 25724911 · PMCID: PMC4424408

96. **PeptideShaker enables reanalysis of MS-derived proteomics data sets**
Marc Vaudel, Julia M Burkhart, René P Zahedi, Eystein Oveland, Frode S Berven, Albert Sickmann, Lennart Martens, Harald Barsnes
Nature Biotechnology (2015-01) https://doi.org/ggkds8
DOI: 10.1038/nbt.3109 · PMID: 25574629

97. **PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data**
Yehia Mokhtar Farag, Carlos Horro, Marc Vaudel, Harald Barsnes
Journal of Proteome Research (2021-10-28) https://doi.org/gpdd85
DOI: 10.1021/acs.jproteome.1c00678 · PMID: 34709836 · PMCID: PMC8650087

98. **mzML—a Community Standard for Mass Spectrometry Data**
Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpp, Steffen Neumann, Angel D Pizarro, … Eric W Deutsch
Molecular & Cellular Proteomics (2011-01) https://doi.org/dxkg99
DOI: 10.1074/mcp.r110.000133 · PMID: 20716697 · PMCID: PMC3013463

99. **File Formats Commonly Used in Mass Spectrometry Proteomics**
Eric W Deutsch
Molecular & Cellular Proteomics (2012-12) https://doi.org/ggkdvv
DOI: 10.1074/mcp.r112.019695 · PMID: 22956731 · PMCID: PMC3518119

100. **Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows**
Kenneth Verheggen, Helge Ræder, Frode S Berven, Lennart Martens, Harald Barsnes, Marc Vaudel
Mass Spectrometry Reviews (2020-05) https://doi.org/gbwkmf
DOI: 10.1002/mas.21543 · PMID: 28902424

101. **Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases**
Lukas Käll, John D Storey, Michael J MacCoss, William Stafford Noble
Journal of Proteome Research (2008-01) https://doi.org/fbxhxp
DOI: 10.1021/pr700600n · PMID: 18067246

102. **False Discovery Rate Estimation in Proteomics**
Suruchi Aggarwal, Amit Kumar Yadav
Methods in Molecular Biology (2016) https://doi.org/f79mzp

DOI: 10.1007/978-1-4939-3106-4_7 · PMID: 26519173

103. **Unbiased False Discovery Rate Estimation for Shotgun Proteomics Based on the Target-Decoy Approach**
Lev I Levitsky, Mark V Ivanov, Anna A Lobas, Mikhail V Gorshkov
Journal of Proteome Research (2016-12-13) https://doi.org/gqtfbj
DOI: 10.1021/acs.jproteome.6b00144 · PMID: 27959540

104. **Quality Control in Proteomics**
PROTEOMICS
Proteomics (2011-03) https://doi.org/bzmh75
DOI: 10.1002/pmic.201190020 · PMID: 21374817

105. **Quality control in LC-MS/MS**
Thomas Köcher, Peter Pichler, Remco Swart, Karl Mechtler
PROTEOMICS (2011-02-07) https://doi.org/c54ck5
DOI: 10.1002/pmic.201000578 · PMID: 21360669

106. **Quality control in mass spectrometry-based proteomics**
Wout Bittremieux, David L Tabb, Francis Impens, An Staes, Evy Timmerman, Lennart Martens, Kris Laukens
Mass Spectrometry Reviews (2017-09-07) https://doi.org/gbs3vs
DOI: 10.1002/mas.21544 · PMID: 28802010

107. **AlphaPept, a modern and open framework for MS-based proteomics**
Maximilian T Strauss, Isabell Bludau, Wen-Feng Zeng, Eugenia Voytik, Constantin Ammar, Julia Schessner, Rajesh Ilango, Michelle Gill, Florian Meier, Sander Willems, Matthias Mann
Cold Spring Harbor Laboratory (2021-07-26) https://doi.org/gqmt2q
DOI: 10.1101/2021.07.23.453379

108. **rawDiag: An R Package Supporting Rational LC–MS Method Optimization for Bottom-up Proteomics**
Christian Trachsel, Christian Panse, Tobias Kockmann, Witold E Wolski, Jonas Grossmann, Ralph Schlapbach
Journal of Proteome Research (2018-07-06) https://doi.org/gd39tz
DOI: 10.1021/acs.jproteome.8b00173 · PMID: 29978702

109. **The rawrr R Package: Direct Access to Orbitrap Data and Beyond**
Tobias Kockmann, Christian Panse
Journal of Proteome Research (2021-03-09) https://doi.org/gjgdxj
DOI: 10.1021/acs.jproteome.0c00866 · PMID: 33686856

110. **RawBeans: A Simple, Vendor-Independent, Raw-Data Quality-Control Tool**
David Morgenstern, Rotem Barzilay, Yishai Levin
Journal of Proteome Research (2021-03-04) https://doi.org/gmwh5f
DOI: 10.1021/acs.jproteome.0c00956 · PMID: 33657803 · PMCID: PMC8041395

111. **SIMPATIQCO: A Server-Based Software Suite Which Facilitates Monitoring the Time Course of LC–MS Performance Metrics on Orbitrap Instruments**
Peter Pichler, Michael Mazanek, Frederico Dusberger, Lisa Weilnböck, Christian G Huber, Christoph Stingl, Theo M Luider, Werner L Straube, Thomas Köcher, Karl Mechtler
Journal of Proteome Research (2012-10-22) https://doi.org/f3sk5j
DOI: 10.1021/pr300163u · PMID: 23088386 · PMCID: PMC3558011

112. **Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data**

Bryan A Stanfill, Ernesto S Nakayasu, Lisa M Bramer, Allison M Thompson, Charles K Ansong, Therese R Clauss, Marina A Gritsenko, Matthew E Monroe, Ronald J Moore, Daniel J Orton, … Thomas O Metz
Molecular & Cellular Proteomics (2018-09) https://doi.org/gd7n2k
DOI: 10.1074/mcp.ra118.000648 · PMID: 29666158 · PMCID: PMC6126382

113. **SprayQc: A Real-Time LC–MS/MS Quality Monitoring System To Maximize Uptime Using Off the Shelf Components**
Richard A Scheltema, Matthias Mann
Journal of Proteome Research (2012-05-11) https://doi.org/gqmt2m
DOI: 10.1021/pr201219e · PMID: 22515319

114. **Metriculator: quality assessment for mass spectrometry-based proteomics**
RM Taylor, J Dance, RJ Taylor, JT Prince
Bioinformatics (2013-09-02) https://doi.org/f5gpzz
DOI: 10.1093/bioinformatics/btt510 · PMID: 24002108

115. **OpenMS: a flexible open-source software platform for mass spectrometry data analysis**
Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, … Oliver Kohlbacher
Nature Methods (2016-08-30) https://doi.org/f82r32
DOI: 10.1038/nmeth.3959 · PMID: 27575624

116. **MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments**
Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, Olga Vitek
Bioinformatics (2014-05-02) https://doi.org/f6j737
DOI: 10.1093/bioinformatics/btu305 · PMID: 24794931

117. **MSstatsQC: Longitudinal System Suitability Monitoring and Quality Control for Targeted Proteomic Experiments**
Eralp Dogu, Sara Mohammad-Taheri, Susan E Abbatiello, Michael S Bereman, Brendan MacLean, Birgit Schilling, Olga Vitek
Molecular & Cellular Proteomics (2017-07) https://doi.org/gbmgrh
DOI: 10.1074/mcp.m116.064774 · PMID: 28483925 · PMCID: PMC5500765

118. **Proteomics Quality Control: Quality Control Software for MaxQuant Results**
Chris Bielow, Guido Mastrobuoni, Stefan Kempa
Journal of Proteome Research (2015-12-28) https://doi.org/f8c54w
DOI: 10.1021/acs.jproteome.5b00780 · PMID: 26653327

119. **protti: an R package for comprehensive data analysis of peptide- and protein-centric bottom-up proteomics data**
Jan-Philipp Quast, Dina Schuster, Paola Picotti
Bioinformatics Advances (2021-12-10) https://doi.org/gqmt2n
DOI: 10.1093/bioadv/vbab041

120. **The NCBI Eukaryotic Genome Annotation Pipeline**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

121. **UniProt** https://www.uniprot.org/help/biocuration

122. https://www.psidev.info/sites/default/files/2018-03/MIAPE_MSI_1.1.pdf

123. **UniProt** https://www.uniprot.org/help/sequence_origin

124. **UniProt** https://www.uniprot.org/help/manual_curation

125. **fasta_utilities**
Phillip Wilmarth
(2022-11-24) https://github.com/pwilmart/fasta_utilities

126. **NCBI Datasets**
NCBI
https://www.ncbi.nlm.nih.gov/datasets/

127. **Eukaryotic genomes annotated at NCBI**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/

128. **Prokaryotic RefSeq Genomes** https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

129. **Eukaryotic RefSeq Genome Annotation Status**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/status/

130. **Training & Tutorials - Site Guide - NCBI** https://www.ncbi.nlm.nih.gov/guide/training-tutorials/

131. **Ensembl genome browser 108** https://uswest.ensembl.org/index.html

132. **Ensembl Genomes** http://ensemblgenomes.org/

133. **cRAP protein sequences** https://www.thegpm.org/crap/

134. **Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics**
Ashley M Frankenfield, Jiawei Ni, Mustafa Ahmed, Ling Hao
Cold Spring Harbor Laboratory (2022-04-28) https://doi.org/gp4xcp
DOI: 10.1101/2022.04.27.489766

135. **Proteomic analyses using an accurate mass and time tag strategy**
Ljiljana Paša-Tolić, Christophe Masselon, Richard C Barry, Yufeng Shen, Richard D Smith
BioTechniques (2004-10) https://doi.org/gqcns4
DOI: 10.2144/04374rv01 · PMID: 15517975