Designing biomedical proteomics experiments: state-of-the-art and future perspectives

Evelyne Maes, Pieter Kelchtermans, Wout Bittremieux, Kurt de Grave, Sven Degroeve, Jef Hooyberghs, Inge Mertens, Geert Baggerman, Jan Ramon, Kris Laukens, et al.

Designing biomedical proteomics experiments: state-of-the-art and future perspectives

Evelyne Maes[1], Pieter Kelchtermans[2], Wout Bittremieux[3], Kurt De Grave[4], Sven Degroeve[2], Jef Hooyberghs[5], Inge Mertens[1], Geert Baggerman[1], Jan Ramon[6], Kris Laukens[3], Lennart Martens[2], Dirk Valkenborg[7]

Abstract

With the current expanded technical capabilities to perform mass spectrometry-based biomedical proteomics experiments, an improved focus on the design of experiments is crucial. As it is clear that ignoring the importance of a good design leads to an unprecedented rate of false discoveries which would poison our results, more and more tools are developed to help researchers designing proteomic experiments. In this review, we apply statistical thinking to go through the entire proteomics workflow for biomarker discovery and validation and relate the considerations that should be made at the level of hypothesis building, technology selection, experimental design and the optimization of the experimental parameters.

1 Applied Bio & molecular systems, VITO, Mol, Belgium; University of Antwerp, Antwerp, Belgium,
2 Med. Biotech Center, VIB Ghent, Belgium; University of Ghent, Belgium
3 Dept. of Mathematics & computer science, University of Antwerp, Belgium; University hospitals Antwerp, Belgium
4 Dept. of Computer Science, KULeuven, Leuven, Belgium
5 Applied Bio & molecular systems, VITO, Mol, Belgium
6 Dept. of Computer Science, KULeuven, Leuven, Belgium; MAGNET, INRIA, Lille, France
7 Applied Bio & molecular systems, VITO, Mol, Belgium; Univerity of Antwerp, Belgium; University of Hasselt, Belgium

Designing biomedical proteomics experiments: state-of-the-art and future perspectives

Abstract

With the current expanded technical capabilities to perform mass spectrometry-based biomedical proteomics experiments, an improved focus on the design of experiments is crucial. As it is clear that ignoring the importance of a good design leads to an unprecedented rate of false discoveries which would poison our results, more and more tools are developed to help researchers designing proteomic experiments. In this review, we apply statistical thinking to go through the entire proteomics workflow for biomarker discovery and validation and relate the considerations that should be made at the level of hypothesis building, technology selection, experimental design and the optimization of the experimental parameters.

1. INTRODUCTION

Since the introduction of 'omics' technologies in biomedical research, a profound change in addressing biomedical research questions has taken place as they might often be preferred over traditional biochemical procedures. As these omics platforms gained popularity with the development of high-throughput measurement methodologies, they can now be seen as central players in scientific research. With mass spectrometry-based proteomics as one of the most mature omics platforms, major advances in the biomedical field have already been achieved. However, despite the recent developments in the field of both mass spectrometry and bioinformatics which led to advanced sensitivity and specificity combined with increased throughput, the translation of results from biomedical projects to clinical practice remains poor. It is therefore clear that the general trend where initially promising findings do not result in improvements in healthcare needs to be halted and that a major focus on both the quality and the design of experiments is mandatory for every new research question. The importance of experimental design however is not new. Over eighty years ago, Ronald A. Fisher raised the bar on the quality of empirical science when he published the eponymous book *The Design of Experiments* [1]. Today, these ideas remain as relevant as ever. Our vastly expanded technical capabilities to perform experiments have made it more, not less important to heed the principle that experimentation requires premeditation. If ignored, we would poison our knowledge by an

unprecedented rate of false discoveries (type I errors) or needlessly slow down our acquisition of new knowledge (through type II errors or wasted experimental resources).

A publication in The Lancet by Petricoin *et al*. still remains one of the best examples to indicate that, when experimental design and subsequent analysis of the results is lacking, a cancer diagnostic assay based on SELDI-TOF experiments can be invalidated [2]. Indeed, although the publication suggested that a near perfect sensitivity and specificity could be achieved to diagnose ovarian cancer (even at an early stage) in serum samples, other groups indicated the presence of confounding factors and lack of reproducibility in both the initial dataset and two related experiments published by the same research group [3,4]. Among the likely culprits were selecting peaks predominantly at low m/z (where the noise from saturation, miscalibration, and matrix molecules was the highest), inconsistent baseline correction, invalidated changes in equipment and protocol, miscalibration of time-of-flight to m/z conversion, possible lack of blinding, lack of randomization, no analysis of variation, and no replication over different values for important variables that cannot be controlled. This finding of severe deficiencies in the experimental design received a reply from the authors of the flawed study, which was itself subsequently thoroughly rebutted [5] . Since then, several other publications have illustrated the importance of experimental design and the consequences of ignoring the issues that must be addressed in order to perform successful proteomic studies [6].

In the design of experiments, at least when non-deterministic processes are considered as in biology and mass spectrometry, statistics plays a crucial role. A first step is to understand what the inputs and outputs to the process are, what their range is, and what input parameters one can control — the control variables. Second, in order to be able to perform a sound statistical analysis after the experimentation is done, it is important to first understand the process that has generated the data to falsify the research hypothesis. In particular, one should analyze which assumptions can and cannot be made with respect to the independence of variables, and what the distributions are from which the observations are drawn. The instrument's maintenance status [7], instrument settings, the ambient temperature, and the perfume of choice of the lab technician can all influence the results. Some of those variables may affect the results more strongly when a particular combination of multiple variables occurs — an effect that requires many more sample runs if we wish to be able to discover it rather than the simple main effects of single variables.

In a longitudinal study (multiple measurements over time) or case-control study (comparison between two populations or treatment groups), all the measurements should be generated by exactly the same protocol. Apart from the variables that were allowed to vary (i.e., time, treatment, disease status), all other parameters need to be controlled such that the measurement can be assumed to originate from the same process. Even though such basic statistical issues seem straightforward, there are a surprising number of studies that do not apply them rigorously. Even worse, it is not always feasible to create the ideal experiments due to various constraints, in which case approximations are needed, which might require custom analyzing and reporting. For example, decaying instrument performance, batch effects, or time effects are often difficult to control and should be dealt with by randomization schemes. In this manuscript we apply statistical thinking to go through the entire proteomics workflow for biomarker discovery and validation and relate the considerations that should be made at the level of hypothesis building, technology selection, experimental design and the optimization of the experimental parameters. An overview of all these steps in the proteomics workflow is also depicted in Figure 1.

A topic that will not be tackled in this review are the strategies to analyze the vast amounts of mass spectral data. The appropriate data analysis strategy is often dictated by the type of mass spectrometry experiment, the objectives of the study and the organism that is scrutinized. For example, data that originates from selected reaction monitoring requires a completely different approach than data from shotgun proteomics.

Most bioinformatics tools for mass spectrometry can be categorized in two classes. A first class of tools deals with identification of the biomolecules from the full scan and fragment ion spectra. The second class deals with the quantitative information and summarization towards the level of the protein, peptide or metabolite. Of course advances in one class like, e.g., *de novo* search algorithms or algorithms that are invariant for post-translational modification, will influence how the second class of tools will incorporate new information in their quantitative analysis. Moreover, it seems that the protein inference problem [8] can be regarded as a special case of the protein quantification problem [9]. Bioinformatics tools for mass spectrometry data are continuously improved [10,11,12] and concepts are lively debated by the scientific community [13].

2. RESEARCH QUESTION AND THE FORMAL HYPOTHESIS

The goal of designing an experiment is to obtain precise and accurate information to answer a specific research question in a most optimal and unbiased way. Hence, the first step in planning an experiment is to formulate the scientific question to be answered. For example, an experiment can be exploratory in nature in a qualitative or quantitative fashion, e.g., which proteins are present in the sample, or which proteins are expressed differently when comparing the protein profile of two groups of patients. These high-throughput discovery studies are often used to generate new lists of hypothetical biomarkers for diagnostics, prognosis or predictive medicine. Other research questions may need a more targeted approach as is the case in verification studies to evaluate potential biomarkers or quantify the presence of a compound in a particular matrix. Hence, these targeted measurements aim to quantify or detect the presence of a protein without comparing the observations with a reference measurement, as is the case in the discovery approach. Moreover, discovery experiments are optimized to identify and quantify the largest possible number of biomolecules from a single sample. To accomplish this objective, the requirements of speed and reproducibility are often tempered. By contrast, targeted experiments are conceived to measure only a few biomolecules, but in a robust, reproducible, and fast manner. Therefore, in discovery experiments the number of proteins that are measured (P), is usually much larger than the number of samples (N), i.e., $P>>N$. For targeted measurements, we have the opposite, $P<<N$.

Once a research question is settled upon, it needs to be formalized into a proper hypothesis. Like most analytical techniques, mass spectrometry-based proteomics is affected by noise and variability. Therefore, statistics is needed to evaluate the meaning of a certain measurement and whether it is significantly larger or smaller than another measure or fixed threshold value. As such, false conclusions from the data or so called chance findings can be controlled when thorough understanding of the noise source and their statistical properties is present. Consequently, it is good practice to formalize your research questions into a falsifiable hypothesis. The research question from the previous paragraph can be framed in a simple hypothesis that examines whether there is a relationship between two variables (e.g., protein expression and treatment effect). This hypothesis can be seen as comparative in nature as it aims at evaluating the difference between protein expression from two treatment groups, or evaluate whether a protein abundance is larger than a pre-specified threshold value.

The simplest comparative designs attempt to gather sufficient evidence to support a single hypothesis H_1. Collecting evidence to directly support a hypothesis has long been recognized as unscientific, as history has shown that there is no shortage of false hypotheses for which copious amounts of supporting evidence can be produced one way or another. The safe way to give credence to a hypothesis is to disprove its alternative, which Fisher called the null hypothesis, H_0. There are two ways a comparative experiment for verifying a single hypothesis can give false results: either H_0 is incorrectly rejected, leading one to mistakenly accept the study hypothesis H_1, or one fails to reject H_0 even though H_1 is true. The former event is a false positive or type I error, while the latter is a false negative or type II error. The significance level of an experiment is the probability of type I errors when H_0 holds, and the power of an experiment is the probability of avoiding a type II error when H_1 holds. While it is trivial to reduce either probability by choosing *a priori* a corresponding significance level at which to reject the null hypothesis, a careful experimental design isn't enough to avoid both types of error: one also needs to run a sufficient number of replicates to understand sources of variation in the data, both from biological and from technological sources. It is evident that these noise sources will influence the shape of the null distribution and consequently will impact the choice of the statistical test and its outcome. Insight into the technological and biological variability can moreover be used to perform *a priori* power calculations that yield the minimal sample size required to reject the null hypothesis with enough "power" when a difference in protein expression is present (BOX 1).

A rule of thumb is to keep the hypothesis simple when the technology and experimentation to falsify the hypothesis become more complex. On the other hand, many scientists want to get the most out of their expensive proteomics experiments and often combine multiple research questions in one experiment. For example, to test the effect of an antibiotic on multiple strains of an organism with regards to protein expression over different time points, whilst focusing on different matrices that require different sample preparation protocols (e.g., cytosolic, and membrane proteins). Data from complex experiments are more difficult to evaluate with simple statistics such as Student's t-test, F-test, Wilcoxon signed rank test, or Mann-Whitney U test. Instead, these complex experiments require more advanced statistical models such as ANOVA, mixed models, or regression models. In these statistical models each question enters as a factor and it is important to carefully design such experiments in order to avoid nested factors. Nesting can be avoided by crossing the factors such that every category of one variable co-occurs in the design with every category of the other variables. In other words, each level of the

factor has a fundamental property that is the same for every level of the other factors in the experiment [14]. Furthermore, crossed experimental designs can test for interactions. In nested designs, some factors cannot participate at all levels and studying interactions is thus not possible. However, in some occasions it is not possible to guarantee fully crossed experimental designs, which has an implication on the choice of statistical model to study the nested data.
After formulating a biological question and translating it into a falsifiable hypothesis, the task is to select the optimal technology and experimental technique to answer it [15]. For a given biomedical question, one should identify the challenges involved and which proteomic approaches are able to address these. Indeed, a variety of proteomic techniques are available that all differ in their versatility, difficulty and overall costs: from global protein expression profiling studies to very focused approaches that measure only pre-specified targets [15].

3. SELECTION OF TECHNOLOGY

Out of all available techniques and instruments, we will focus the rest of this review on protein discovery and targeted protein identification with LC-(MS)n. Both techniques start from digested, potentially labeled, protein samples. While a reproducible sample preparation is essential, optimal protein extraction from different samples is still approached purely empirical [16,17,18,19] and few computational resources exist to improve this step.

3.1 Considerations at the wet-lab level

3.1.1 Digestion

In peptide-centric proteomics approaches the proteolysis of proteins into MS-detectable peptides is a crucial factor to achieve decent protein and proteome coverage. However, in many cases protein sequence coverage is low, with some proteins identified by only a single peptide sequence. In a typical experiment, trypsin is used as protease and cleaves at the C-terminus of lysine and arginine, producing ‘predictable’ peptides of practical length and favorable charge which are well suited for fragmentation by collision-induced dissociation (CID) [20]. Of the theoretically possible tryptic peptides, the ones that are typically identified are from 7 to 35 amino acids long, and carry at least a positive charge at each terminus, yielding mostly doubly or triply charged peptides. However, the use of a single protease will often not suffice to provide full sequence coverage, and the use of alternative proteases can therefore expand coverage

[21]. As a result, orthogonal digestion is increasingly popular, with endoproteinase Lys-C, cleaving at the C-terminus of lysine, and chymotrypsin, which cleaves at the C-terminus of phenylalanine, tyrosine, tryptophan, and leucine as the most popular alternative proteases. Due to a less predictable ionization and fragmentation pattern, the peptides obtained from these proteases can be more difficult to identify. Specific cases where orthogonal digestion has proven to be beneficial are the study of post-translational modifications, of highly folded proteins, and of splice variants [22]. In extreme cases, even more than two proteases can be used, as shown by Swaney et al, where a *S.Cerevisiae* proteome digestion by five proteases (Trypsin, Lys-C, Arg-C, Glu-C and Asp-N), led to a doubling of the sequence coverage [23]. A comparative table of theoretical whole-proteome coverage for *Homo sapiens* based on various parallel enzymatic digests can moreover be found in Vandermarliere *et al*. [20].

3.1.2 Fractionation

The most common problem in proteomics is that biological samples are too complex to result in discernible features after LC and MS separation. Because of this, prefractionation of samples is beneficial if sufficient material is available, though there have been attempts to analyze samples without pre-fractionation [24,25]. A large variety of fractionation methods exist, including multidimensional chromatographic separations (e.g., reversed phase, ion exchange, size exclusion, and hydrophobic interactions) which generate separate fractions, SDS gel-based fractionations or those that selectively obtain a particular subset of proteins or peptides (e.g. affinity-based beads) [26]. With this wide variety of different fractionation strategies, choosing the optimal methods might be difficult. Indeed, as not all methods are equally suitable for the samples of choice, a good consideration of the fractionation methods remains crucial [27]. Even more important, if a preferred prefractionation method is performed, in a comparative biomedical study it must similarly be performed for all samples. Additionally, it should be mentioned that fractionation will introduce additional variability in the sample, which makes good reproducibility and minimal instrument drift crucial, especially when performing quantitative experiments. However, this induced variability can be accounted for by an experimental design that employs isotope label for the relative quantitation of proteins and peptides, e.g., isobaric labels. Doing so, samples are pooled and fractionated simultaneously such that all the obtained fractions are affected by the same technical variability, making them directly comparable in a statistical analysis.

3.1.3 Co-purification: Pull-down / depletion / enrichment

For several research questions it might be necessary to perform a depletion or a pull-down of a certain subtype or class of proteins [28]. A wide variety of strategies exist, and depend on the biomedical tissue or protein class of interest. One of the most typical examples are biomedical projects where (human) serum/plasma samples are frequently used to deplete highly abundant proteins in order to decrease the dynamic range and thus to increase coverage of lower abundant proteins. Several different tools exist to achieve the depletion of most of the highly abundant proteins, including Multiple Affinity Removal System (MARS) and ProteoPrep [29]. Alternatively, other techniques which aim to limit the extent of the dynamic range in blood samples by enriching the proteins of low abundance, e.g., ProteoMiner beads, can be used as well. However, it is important to note that, although empty beads are of regularly use in Western analysis, empty beads are of limited use in proteomics as they should not be used to quantify unspecific binding as the bound protein population is always vastly different from the actual pulldown sample. Some other research questions however, require the pull-down of specific proteins or a special class of proteins. For example, when a strong focus is present on plasma membrane proteins, the acute slice biotinylation assay (ASBA) technique can be used [30]. In designing these kind of experiments, it is crucial to keep in mind that expression profiles from extreme experimental conditions (e.g., controls are pull-downs, cases are not) should be avoided, especially since depletion or enrichment tends to be transitive [31]. Additionally, one should also take into account that many of these depletion technologies will also eliminate associated proteins (such as albumin binding proteins, antibody bound antigens, etc). Furthermore, one should be aware of the implicit assumptions made by the normalization method used, as these often do not hold for co-purification experiments.

3.2 Considerations on the type of experiment

LC-MS-based experiments can roughly be divided into two subcategories: discovery (untargeted) MS, and targeted MS. The first type of experiments are more comparative, while the second type of experiments are more quantitative. The major principles behind these two LC-MS-based categories in proteomics approaches are summarized in [32,33,34,35]. As the name implies, protein discovery experiments aim to identify as many proteins as possible in a sample of unknown content. In these workflows, protein samples are digested into peptides and fractionation of the resulting peptide mixture takes place before it is subjected to mass

spectrometric analysis. Despite the ability to identify and quantify thousands of proteins in complex samples [18,36], it still remains extremely difficult and time-consuming to reveal the composition of a complete, complex proteome.

For discovery based methods, an important methodological decision that has to be made is whether the MS acquisition will be performed in a data-dependent or data-independent way. Measuring in data-dependent acquisition (DDA) mode, where precursor ions are selected automatically from those detected in a survey scan immediately preceding the ion selection, can result in incomplete sampling of complex peptide mixtures. Indeed, in DDA the number of peptides that are sampled within one LC-MS/MS run is limited by the local sample complexity and concurrent MS/MS speed. This is detrimental for low abundant peptides in complex samples as a single MS spectrum might contain over 100 molecular entities, of which only a handful can be selected for fragmentation before the next full scan [37]. As a result of this bias towards peptides that yield high intensity signals, reproducible quantification of more stochastically sampled low abundance peptides remains challenging.

Strategies which do not require the detection and/or knowledge of the precursor to trigger acquisition of fragment ion spectra are categorized as data independent acquisition (DIA) experiments. In data-independent acquisition, all peptides present in a defined mass-to-charge window (known as a bin), are subjected to simultaneous fragmentation. As such the link between precursor and fragment ions is lost which complicates the analysis of the resulting data sets, and scan performance can be directly impacted by the isolation window width. Therefore DIA technology depends on a trade-off between the number of bins and the bin width, and even more crucially, on sophisticated, tailor-made bioinformatics for the identification and quantification of the peptides. This reliance of DIA approaches on dedicated bioinformatics solutions is due to the wide isolation window (usually >10Da) that intentionally leads to the co-isolation and co-fragmentation of multiple peptides, which in turn results in much more complex fragmentation spectra. On the other hand, smart strategies like multiplexed DIA [38] exist that combine a high scan performance with narrow isolation windows. However, demutliplexing of the obtained data require even more advance bioinformatics methods. A special type of DIA is the SWATH-MS methodology trademarked by SCIEX that provide high resolution fragment ion spectra of all precursors in a user-defined precursor ion window [39]. In SWATH-MS, a DIA acquisition method is combined with targeted data analysis. Typically, precursor ions from bins of 25 m/z units are fragmented, recorded with a time-of-flight mass analyzer and extracted ion

chromatograms of peptides of interest are generated [40]. The method thus allows to quantify as many components as identified with typical DDA experiments, but with the accuracy of SRM.

Quantitative DDA proteomics experiments can be performed using either relative or absolute quantification [41]. Although label-free shotgun experiments are mostly applied when peptide expression levels in a large cohort of samples need to be compared on a relative scale, the multiplexing capability of the labeled workflow is often preferred when limited sample material is available and sample preparation is costly. These 'multiplexing labeling procedures' offer a broad variety of approaches to choose from, such as tandem mass tags (TMT) [42], isobaric tags for relative and absolute quantification (iTRAQ) [43], and stable isotope labeling with amino acids in cell culture (SILAC) [44]. Absolute quantification is typically carried out based on relative quantification of sample peptides against spiked-in synthetic peptides labeled with heavy isotopes. Approaches for shotgun proteomics experiments include the absolute quantification (AQUA) [45] strategy or QconQAT [46]. To select proper peptide candidates, computational packages can filter an *in silico* digest of the protein of interest for those peptides that are unique across the organism, detectable with MS, and preferably likely to remain unmodified. Of course, it is also important to take note of the limitations of these various techniques when selecting the most suitable DDA approach for the study at hand [47]. For example, although multiplexing strategies have many advantages, it remains important to fully understand the ratio compression due to co-isolated and co-fragmented ions and the consequences it has on the obtained data [48,49].

In contrast to discovery experiments, targeted proteomics experiments are based on an *a priori* defined set of proteins of interest. This set can be determined from prior biological knowledge or from previous discovery experiments. The sample processing protocol is essentially identical to that of discovery approaches, in that proteins are first digested to peptides, which are then separated with liquid chromatography and analyzed with MS. The major difference is found within the mass spectrometer, where peptides of interest are now targeted for fragmentation, followed by quantitative detection of one or more of the resulting fragments. This enables the isolation (or can even disprove the presence) of one or more proteins of interest in a sample with very high specificity and sensitivity [50]. The combination of a precursor mass and one of its fragment ion masses is referred to as a transition. Two different targeted approaches exist: selected reaction monitoring (SRM; sometimes also referred to as multiple reaction monitoring or MRM), and parallel reaction monitoring (PRM). In SRM, only a subset of transitions are

typically monitored. This way, SRM experiments filter out background signals, which increases the signal-to-noise ratio and thus achieves a more accurate quantification when compared to untargeted experiments [51]. Although SRM measurements are still the gold standard for targeted analyses, the development of high resolution/accurate mass (HR/AM) instruments is revolutionizing the field and has enabled the development of the PRM technique as an alternative targeted approach [33]. PRM differs from SRM in that it monitors multiple MS/MS fragmentation channels simultaneously, which not only allows more precise quantification of peptides, but also enables further improved detection limits. PRM relies on highly accurate mass measurements of both precursor and fragment ions, which allows for a narrow precursor ion selection window.

3.3 Considerations on the technology

A technology choice also implicitly enforces method-related constraints that need to be considered, as these parameters cannot change and might limit the scope of the experiment. These parameters are found for both the LC as well as the MS technologies.

3.3.1 Mass spectrometry

3.3.1.1 Mass accuracy

Generally defined as the measured mass minus calculated mass in parts per million (ppm), mass accuracy is a crucial parameter to take into account when answering certain biomedical research questions. Not only will it influence the false discovery rate of protein and peptide identification, as a molecular mass of a peptide determined with 1 ppm mass accuracy rules out about 99% of amino acid compositions possible for a given integer or nominal mass [52], it also influences the ability to distinguish isotopic patterns. This is important in peptide intensity- based quantification and in charge state determination.

3.3.1.2 Mass resolution and mass resolving power

The mass resolution of an MS instrument influences whether a mass spectrometer can unambiguously determine the elemental composition for confirmation of empirical formulas or identification of unknowns. The better the mass resolution, the less likely it is that a mass peak of interest will be merged with an interfering ion from the sample or background. Mass resolution can thus be defined as the minimal mass difference between two spectral peaks

which can be clearly distinguished. Measuring with higher mass resolution requires more measuring time, and a trade-off between resolution and time should therefore be made. Typically, high resolving power can be found in orbitrap instruments (100.000-240.000 Full width at half maximum (FWHM) at a given m/z), followed by time-of-flight instruments (10.000-60.000 FWHM at a given m/z), while quadrupoles and especially ion traps have lower mass resolutions (1.000-20.000 FWHM at a given m/z).

3.3.1.3 Scan duty cycle

Duty cycles are most often defined as the proportion of time during which a system is operated; for mass spectrometers, the typical meaning of duty cycle refers to the amount of time spent on an MS or MS/MS scan. It is important to consider how to use this scan time optimally: how many MS/MS spectra *per* full MS do you want to obtain, what target resolution do you want to achieve in the MS1 and MS2 levels, and do you want different fragmentation times? Because of the trade-off between resolution in MS1 and the number of identifications in MS2, optimization of the duty cycle needs to be carefully considered in light of the experiment at hand.

3.3.1.4 Dynamic exclusion

Most data-dependent or targeted selected ion monitoring (SIM) measurements employ the dynamic exclusion feature, in which a mass is excluded over a defined period of time (typically ranging from 15 – 600 s) after it is analyzed. This provides the instrument with a better chance to analyze less abundant ions that are eluting at the same time as the more dominant but excluded ions. Indeed, implementing dynamic exclusion is very beneficial for biomedical samples as the complexity of these samples can easily overwhelm the separation efficiency of LC instruments. Without dynamic exclusion, highly abundant peptides will be repeatedly selected for fragmentation in data-dependent measurements, at the cost of overall proteome coverage [53]. However, in low complex samples, it might be beneficial to disable dynamic exclusion such that multiple scans from a data dependent acquisition can be employed to increase spectrum quality by selecting the fragment spectrum at the apex of the chromatographic profile such that ion abundance is most optimal for fragmentation. Alternatively, multiple fragment spectra corresponding to the same chromatographic profile can be merged to improve spectral quality.

3.3.1.5 Mass range / mass window

Another factor that should be taken into account is the required mass range. The mass range is the range of m/z's over which a mass analyzer can operate to record a mass spectrum. Depending on the type of molecules (e.g., lipids or proteins) different windows are needed, as different mass ranges are required to obtain a sufficiently broad overview. In addition, depending on the type of mass spectrum (e.g. full spectrum, fragmentation spectrum), different settings are applied. The width of mass isolation windows used for precursor ion selection is another crucial factor that needs to be determined. Mass windows that are too small (e.g., 1 m/z) lose essential information about isotopes, while setting up too large a window can lead to the inclusion of co-eluting molecules. Additionally, the type of instrument is also a crucial factor in deciding the width of the mass window or the instrument resolution, as in quadrupoles for example, setting a mass window too narrow will result in lower transmission and thus lower sensitivity and spectral quality.

3.3.1.6 Spectral accuracy

In some experiments, the isotope pattern can be of interest. In these cases, the spectral accuracy, which indicates the similarity between a measured spectrum (and thus the isotope patterns) and its theoretical spectrum, are of major importance. Furthermore, some software relies on spectral information, such as the relative peak heights in isotope clusters, to aid in the identification of the biomolecule. A user of such software should be aware of the spectral accuracy of the employed instrument because several types of error can compromise the relative peak intensities and corrupt (i.e., add uncertainty to) the molecular identification.

3.3.2 Liquid Chromatography

Mass spectrometry is the central technology in proteomics and is used for detection and identification of thousands of proteins and peptides in a single experiment. However, in order to deal with the high complexity of typical proteomics samples, hyphenated techniques must be used to increase separation power (peak capacity), selectivity, the measured dynamic range, detection sensitivity and sample throughput. In short; a single peak in the mass spectrometer should correspond to a single peptide and should be detected sensitive enough. This is especially true for quantitative analysis in which case interference with other peptides in the analysis needs to be avoided as much as possible. Liquid chromatography is the preferred

hyphenated tool in proteomics research as it provides high-speed, high-resolution and high-sensitivity separation of macromolecules. In chromatography a number of parameters have a big influence on reproducibility and should therefore be controlled.

1) Flowrate and gradient mixing:

Over the years, many different techniques have been developed to increase peak capacity and sensitivity in LCMS and applied to proteomics. The most successful LC technique in proteomics is without any doubt the application of miniaturized (nano) chromatography in which very low flows are used (few 100nl/min) on very narrow columns [54]. However, although nano-LC greatly improves sensitivity of the analysis, it requires miniaturization of pumps, controllers and plumbing in the HPLC which has detrimental effect on reproducibility. Furthermore miniaturization of the LC system also influences the accuracy of several important LC characteristics including control of flowrate, solvent mixing and gradient formation. For this reason, often higher flowrates are chosen for targeted proteomic applications in which high reproducibility is required [55]. In addition higher flowrates have the advantage that ultrafast gradients (UPLC) can be used increasing sample throughput. Higher reproducibility is also achieved when splitter-free HPLC systems are used.

2) Temperature:

Column temperature plays an important role in controlling peak spacing in reversed-phase liquid chromatography separations and temperature has a big influence on retention time. Excellent temperature stability can lead to a high degree of reproducibility, therefore a form of temperature control is essential if one wants to obtain good reproducibility. Indeed, fluctuations in column temperature can give rise to peak retention times drifts, changes in viscosity of mobile phases and overall changes in LC pressures. These changes can have tremendous effects when comparing multiple LC-MS experiments, in particular in label-free experiments, where these changes create biases and difficulties to perform chromatographic alignment of peptide elution peaks [56,57].

3) Back pressure:

Column back pressure is evidently linked directly with flow rate and solvent composition which need to be controlled precisely. However, column back pressure is influenced greatly by column temperature. When temperature increases, the column back pressure decreases, as the temperature decreases, the back pressure increases. This property is useful when working with

columns with small particles (<2 µm) such as those used in UPLC or probably more relevant for proteomics, the use of long (nano-HPLC) columns in which case good temperature control is essential [58].

4. EXPERIMENTAL DESIGN

4.1 Observational studies

In the biomedical field, the vast majority of variables cannot be freely controlled but are conditioned on the individuals included in the study. It is not possible to set variables at the biomolecular level independently to arbitrary levels as the experiment designer is constrained to merely selecting subjects and sometimes their treatment.

In proteomics studies, most observational research methods are based on a cohort study or a case-control study. **Cohort** studies can be defined as studying a group of people with predetermined characteristics who are followed up over a period of time to determine incidence of or mortality from some specific diseases. Cohort studies can be prospective studies where sample collection starts before the individual develops the disease, or retrospective when the samples have been collected in the past for other purposes. **Case-control** studies, on the other hand, determine the relative importance of a variable in relation to the presence or absence of a disease [59]. In general, this type of study will retrospectively compare two (or more) groups. One group has the disease or outcome of interest, and the other group is constructed to consist of people closely matching the first group. Case-control studies are thus able to generate a lot of information from a relatively small group of samples but a decent specification of both the study group and how to construct a valid control group are required.

4.2 Exploratory experiments

In section [3.2], we distinguished discovery and targeted experiments: in MS, it is customary to first cast a wide net to discover molecules potentially correlated with the disease, and only in a later stage measure more carefully a small number of molecules with higher accuracy. If we are not restricted to an observational study, the same principle holds for finding interesting control variables (treatment parameters). The first step of a study typically consists of performing exploratory experiments, also called screening experiments. When there is no significant knowledge available on which factors influence a response of interest, it is conservative to

consider a high number of factors in a first phase. The Pareto principle, also called the 80-20 rule, is the assumption that only a small subset of these effects account for most of the variation seen in the response. It is exactly the identification of these few effects which is of interest in screening experiments. As generally this is only the first step in experimentation, one typically pays attention to performing this step as economically as possible: usually only main effects are investigated, sometimes in combination with two-factor interactions, but even these are confounded in many screening designs. Furthermore, one typically only employs two-level designs to minimize the required number of samples. After identifying the most significant factors, one can proceed to design an experiment to fit a response surface model. For a more elaborate discussion on screening experiments, the reader is referred to [60] and [61].

4.3 Building blocks

Once the variability of the biology and the limits enforced by technology are understood, you can decide on an optimal design to reject the study's hypothesis. A good overview of the principles of experimental design in the context of MS is given in [62] and a more limited introduction to some of the concepts can be found in [63]. A summary of a typical experimental process in proteomics is described in [64]. We will revisit the main principles of experimental design here very concisely. The basic tools of experimental design are: controls, replication, randomization, blocking, and pooling.

4.3.1 Controls

Controls are reliable reference signals with which the signal of interest can be compared. In mass spectrometry they can take the form of regular quality control samples of known composition, or in a comparative study the samples of the untreated group. To achieve a flawless experimental design, it is essential that the control samples undergo the same procedure as the treatment group samples to the maximum extent possible, otherwise an unrecoverable bias may be introduced in the signal.

4.3.2 Replication

Replication is necessary to allow any statistical conclusions. Any signal observed in a proverbial "sample of one" may or may not come around by chance. Only after observing an effect multiple times, it is more convincing that a difference in the signal is likely a true, reproducible difference between two groups. Replication can happen at different steps in the mass spectrometry

pipeline. Biological replicates are samples from separate individuals, or at least from separate locations in the same individual. They are processed independently, but in parallel, over the whole pipeline. Technical replicates are generated from the same biological sample at some point in the pipeline. Biological replicates carry more information as averaging over their measurements allows to cancel out process variation as well as biological variation. Technical replicates are used because they allow for an estimate of how much of the variation is due to the process alone, and because they are cheaper and easier to create. However, in reproducible proteomics platforms, technical variability is mostly negligible compared to biological variability and in all these cases, biological replicates should always be preferred. When samples are subdivided in multiple batches, it is important to note that these samples must be performed in parallel and that independent experiments are difficult to compare.

4.3.3 Randomization

Randomization of the allocation of samples to the available treatment groups and processing slots prevents systematic bias from process drift and other confounding variables that cannot be controlled. Studies without random assignment are quasi-experiments. They cannot prove causalities because the observed effect may as well be caused by a confounding variable. A well-known confounding variable is changing instrument performance over time. Running the treatments groups unmixed, one after another, exposes the later groups to a differently performing instrument than the earlier groups, an effect which may dominate any difference in properties between the groups. A randomized order of samples will approximately balance out such differences.

4.3.4 Blocking

Blocking is the fair allocation of different values of variables that can be controlled to the samples. It is used to prevent bias from the variables that can be controlled and is superior to randomization because the allocation can be balanced between the sample groups, which reduces variation and allows for more accurate estimates of the influence of the variable. Neglect of blocking and randomization can be a fatal flaw in a proteomics experiment, rendering the acquired data worthless and, worse, tempting the experimenter to draw false conclusions, leading to embarrassment [65,5].

4.3.5 Pooling

Pooling is the mixing of multiple samples of the same group as a mean for signal stacking and cancelling of noise. Pooled designs are attractive because they can minimize the experimental cost. In some cases, pooling can be necessity to provide sufficient material for experimentation. Mostly, pooling is done in an effort to reduce the effect of biological variability, since it results in biological averaging for proteins and a lower overall variability. In general, the variance reduction due to the pooling of samples should compensate the reduction in sample size such that the statistical power is maintained. Via statistical theory, optimal pooling designs can be developed that specify the number of samples and experiments to obtain the statistical requirements that are equivalent to a design that does not employ pooling. Often such an equivalent design results in a slight increase of the total number of samples. However, pooling should be used with caution as it makes the interpretation of the outcomes more tenuous. For example, when concentrations vary exponentially, a single high-dose sample can dominate the signal of its whole pool. Pooling should therefore only be used when the effects are linear [66,67,68,69]. Another point of attention are outliers that cannot be detected in pooled designs and often lead to an underestimation of the protein-specific variance.

4.4 Design methods

A number of higher level tools have been developed that combine the previous elements in commonly used patterns for experimental design. We can distinguish fixed strategies and flexible strategies. In fixed strategies, the complete experiment is fixed in advance. Advantages include the simplicity of the post-experiment analysis, the prior knowledge of the relation between the observations (e.g., maximizing the independence of the observations, or the spread over the space of control variables), and the possibility to assess in advance the statistical power towards certain hypothesis tests or the accuracy achievable when estimating the parameters of the model. Flexible strategies perform experiments in sequence, deciding lazily on the next experiment to perform when the previous one is finished or when it is time to decide on the control variables. Its crucial advantage is that attention can be directed to regions in the control variable space which are most interesting or least understood, saving experimental budget (Figure 2).

4.4.1 Fixed design methods

The standard, classical experimental designs are fixed designs. A widely made assumption in classical design is that the biological phenomenon under study varies smoothly under moderate changes of the control variables (factors), so that it is sufficient to consider only a few levels for each factor. The most thorough of classical designs is the full factorial design, in which experimental runs are executed for all combinations of the factor levels. As the number of runs in this approach is exponential in the number of experimental factors, it is only feasible for small design problems. Nonetheless, it has been successfully used, for instance by [70].

To overcome this issue, one can obtain a fractional factorial design by intelligently choosing a subset of the full factorial design. However, this comes at a cost as the smaller the chosen subset, the more the main effects will be confounded with interaction effects. That is, it is impossible for any subsequent data analysis to disentangle effects caused by particular value combinations of two (or more) factors from effects caused by a single factor [71] used this approach to improve proteome coverage on an LTQ-Orbitrap by evaluating the effects of 9 instrument parameters. Both fractional and full factorial designs however have an additional disadvantage, since the number of runs needs to be a power of the considered number of levels.

Plackett-Burman designs are mainly used for screening purposes to estimate main effects as they are typically confounded with the interaction effects. They are very economical as for instance up to 11 two-level factors can be estimated in a 12-run design. An example can be found in [72].

When the goal is to fit a second-order response surface, more apt designs are available, such as the central composite design and the Box-Behnken design. The former starts from a (fractional) factorial design, but it is augmented by including more levels of the experimental factors as this allows to estimate curvature of the response. The latter, by contrast, does not start from a fractional factorial design, typically allowing for a more efficient estimation of the effects.

These classical design approaches can typically be found in statistics handbooks and the designs are easily created, even by hand. The downside is that these approaches are not very flexible; they only allow certain numbers of runs and in case there are additional constraints they

cannot be easily used. Adapting the problem to fit these designs often does not lead to good solutions.

4.4.2 Flexible design methods

Compared to fixed designs, which are normally theory driven, flexible designs allow for more freedom during the data collection process. One advantage is that you can direct attention to regions in the control variable space which are most interesting or least understood. Since the several observations made are not independent, the analysis of experimental results is more complex and fragile. Still, theoretical guarantees w.r.t. the informativeness of the obtained measurements can be provided. E.g., a simple and well-studied setting, known as the 'multi-armed bandit problem' [73], considers a situation where there are several discrete options (e.g., treatments) each producing a randomized output and we want to perform experiments to determine the option giving the highest expected output. In this idealized situation, simple adaptive experiment selection strategies exist that are provably optimal [74]. In general, the use of flexible experimental strategies to fit statistical models is investigated in the field of active (machine) learning [75], while the problem of finding through experiments the best control values to achieve a particular goal is studied in the field of function optimization. Much of the function optimization literature assumes that function evaluations are inexpensive compared to the optimization algorithm itself, which is definitely not the case for MS experiments. Fortunately, since the introduction of Efficient Global Optimization (EGO) by Jones [76] more attention has gone towards optimization using as few function evaluations (sample runs) as possible [77,78,79,80]. This is achieved by fitting a model to the measurements made so far, and choosing the next experiment based on the predictions by the model and an 'infill criterium'. The model works as a surrogate for real experiments during the *in silico* selection of the next experiment. The complexity of the model can vary according to the relative cost of the experiment and the computation. Gaussian processes are the typical choice if the cost of executing an experiment dwarfs any computational cost, but faster alternatives are available [81,82]. Efficient surrogate-based optimization methods assume a continuous design space, but the same methodology also works in discrete spaces, even in the very high-dimensional chemical space [83]. The statistical analysis of the measurements obtained from a flexible strategy is far more delicate, as the samples are not independently and identically distributed.

Depending on the purpose of the experiment being designed, the above tools can be applied in various ways. As pointed out, we can distinguish between explorative experimentation, and experiments where we want to answer specific questions.

4.5 Hypothesis tests

Several strategies exist to determine what experiments to perform. Still, the choice for a particular strategy will depend for a large extent on the goals one wants to reach. Before deciding on a strategy to follow, it is important to analyze the models one wants to build and the hypotheses one wants to test, and hence the expected accuracy of these models or the expected power of these hypothesis tests a proposed strategy may yield. For example, if one wants to test a collection of hypotheses each comparing two groups, it is important to ensure that the selected design will contain of each of the groups compared in each hypothesis a sufficiently large number of cases.

Strategies where experiments are selected as a function of earlier results, such as active learning, may allow to reduce the cost of building a good predictive model for a practical application. However, a disadvantage of such methods is that data obtained in this adaptive way cannot be used easily for hypothesis testing as normally the assumptions underlying such tests are not satisfied.

4.6 Iterative designs

Sometimes, it is hard to estimate in advance how large a sample size is needed to perform the analysis. In such cases, an iterative design may be appropriate. The main idea is to first analyze a smaller sample cohort, and only if the result is unsatisfactory (e.g., the model is insufficiently precise or the null hypothesis can't yet be rejected) one next invests in a larger sample size.

One easy way to do so is to make two satisfactory designs, one of which contains a subset of the experiments of the other one. One can then first perform the smaller design and then reevaluate.

Care needs to be taken when many iterations are performed and evaluated iteratively using hypothesis tests. Some hypothesis tests (e.g. the t-test) are robust, i.e. they still yield

conservative p-values even if performed iteratively on a growing sample, but this is not applicable to all hypothesis tests.

5. OPTIMIZATION OF EXPERIMENTAL PARAMETERS

Once a researcher has settled on the research hypothesis, selected the optimal technology for falsifying the hypothesis, and designed an experiment to control the confounding factors, the experimental parameters can be optimized to generate a data set rich of information. Of course this depends strongly on the selected technology, but a number of models exist that can be applied in the shotgun or SRM/PRM quantification settings. Within one technique it is possible to explore the parameter space for an optimal experimental result [51,84]. While optimizing these experimental parameters, two different approaches are followed for shotgun discovery experiments versus targeted SRM experiments. Since shotgun experiments are interested in detecting as many peptides as possible in the samples, it can be referred to as a bulk process in which parameters are only optimized in a general manner. In targeted analyses, on the other hand, an optimization process must be performed for each a priori known peptide of interest separately, which requires individual optimization of transition-specific parameters in order to achieve maximal signal and sensitivity [85]. Large-scale SRM assays, where hundreds of peptides need to be targeted, require even MS instrument parameters that work well with the broad diversity of peptides to be targeted. Software packages and computational studies attempt to improve or assist determining these parameters for a given experiment [86]. These efforts start with modeling the proteomics pipeline [87] to optimize some experimental parameters. With the latter application fields in mind, we will describe where improvements to the experimental setup can be made when using computational tools.

5.1 Protease activity

In peptide-centric LC-MS methods, the digestion of proteins into peptides (i.e. proteolysis) is an important aspect. Although trypsin can be seen as the standard protease used in most shotgun and targeted approaches, other proteases can be beneficial in the workflow as well. To optimize an experiment, it might thus be beneficial to know which proteins will be cleaved by a certain protease, and in case of a priori known proteins to be measured in targeted experiments, which

peptides will be cleaved from a certain protein. This information is crucial as only these peptides will be analyzed via LC-MS and thus can be detected in the experiment. Predictive proteolysis models can thus improve identification rates in shotgun proteomics and/or provide a priori prediction of suitable peptides for targeted proteomics analyses [86]. Software predicting cleavage probabilities exists for many proteases [86], with as usage mode the theoretical digestion of a single protein or mixture. For trypsin, for example, Cleaving prediction with decision trees (CP-DT) uses the positional information of amino acid sequences around the tryptic site to estimate whether or not the protein will be cleaved [88]. By ranking the peptides by the probability that they will occur after tryptic proteolysis, a list of peptides which can be potentially detected is generated.

5.2 Liquid Chromatography: Retention time prediction

The separation of digested peptides with LC prior to mass spectrometric analysis is one of the most used separation techniques in proteomics. As the amount of time that a peptide is retained on a LC column, i.e. the retention time (RT), is independent of the information present in the MS/MS scan, LC retention time represents another parameter that can be computationally optimized. In shotgun discovery experiments, prediction of the retention time can be used to increase peptide identification confidence [89], while for targeted applications, standard gradients can be customized to make sure the targeted peptides have little overlap with other peptides.

Although the prediction of peptide LC retention times might struggle with the large variety across LC methods and analyses, predictive models have been built for chromatography in proteomics, helping experimentalists by providing expected elution times or hydrophobicity indices (in case of reverse phase chromatography) [86]. While the first RT prediction models assumed that peptide RT is a linear function of the amino acid sequence [90], more recent models also focus on peptide length or positional effects of the amino acid residues [91,92]. Even more sophisticated models, including SSRCal, calculate retention time as a weighted sum of retention coefficients for the individual residues in a peptide and then correct for empirical factors such as length influence and the tendency to form helical structures [93]. To accommodate for different experimental LC conditions Moruz *et al.* proposed to derive a retention index for a specific condition using data driven regression algorithms [94]. There tool ELUDE is fully portable to new chromatographic condition and works for post-translationally modified peptides as well [95].

Besides computational models, also retention time peptide standards can be added to samples to normalize peptide retention time across multiple LC-gradient elution profiles [96].

5.3 Ionization efficiency/ peptide detectability

The analysis of complex digested samples by LC-MS/MS goes hand in hand with the current inability to detect all eluting digested peptides in one LC-MS experiment. Computational methods, however, are able to predict the ionization efficiency of a certain peptide and although detectability is the result of poorly understood processes of getting the peptide in solution subsequently ionized and picked up by the ion detector, it is known that the likelihood to detect a peptide in a certain proteomics experiment depends on four major factors: 1) the chemical properties of the peptide, 2) limitations of the peptide identification protocol, 3) the abundance of the peptide in the sample and 4) the presence of competing peptides in the sample. Kelchtermans *et al.* provide an overview of different computational models which predict the peptide detectability based on these factors [86].

Again, these peptide detectability predictions can be used to address several problems in proteomics experiments. In discovery experiments, peptide detectability can be used to guide protein inference problems and to help label-free quantification [97]. For targeted experiments, the prediction of the detectability of proteotypic peptides is very helpful to optimize the SRM transitions to be detected in an experiment. Here, commonly occurring highly detectable peptides might crowd out the peptides of interest. Based on prior knowledge these very abundant peptides can be identified and the experiment can be set up to minimize their interference. For example, the common Repository of Adventitious Proteins (cRAP, http://www.thegpm.org/crap/) contains a list of contamination proteins that are commonly found in proteomics experiments. Based on their physicochemical properties, the SRM transitions can be optimized to avoid interference by peptides originating from these commonly occurring proteins. Of course, this does not have to be limited to only contaminants, but can include other frequently occurring highly proteotypic peptides that are likely to have a major influence as well.

5.4 Fragmentation modeling

Peptide identification in standard LC-MS/MS-based proteomics experiments typically relies on the prediction of fragment ions and the quality of the experimentally determined fragmentation

spectra for database-driven target identification [98]. In an MS/MS spectrum, the intensity of the peptide fragment ions is dependent on both the abundance of the peptide as well as the efficiency of bond breaking. Although some naïve models make the assumption that all peptide bonds break with equal probability and that each resulting fragment will take on all charges below that of their precursor ion, experimental spectra are more complicated. Predicting the fragmentation patterns of a peptide and the fragment ion intensities is therefore of crucial importance to understand the patterns behind peptide fragmentation [98].

Fragmentation of a peptide bond is either a charge-directed process, which involves a mobile proton migrating to the bond, or a charge-remote process, which is determined by the delicate balance between the total number of available protons and the number of proton sequestration sites (basic amino acids) [98].

Whereas early peptide fragmentation prediction tools such as MassAnalyzer [99] implement a deductive physicochemical model of peptide fragmentation based on this knowledge, current state-of-the-art prediction tools such as PeptideART [100] and MS2PIP [101] employ a fully data-driven machine learning approach to compute accurate peptide fragmentation models from the amino acid properties in a peptide. This information can be beneficial both to increase the protein identification (coverage) in discovery experiments as well as the proteotypicity of the fragment ions for SRM targeted assays.

For targeted SRM experiments, collision energies (CEs) are frequently optimized for every target peptide individually to increase the fragment ion intensities in order to attain the maximum sensitivity. However, instead of optimizing these CEs empirically for each peptide, predicting the optimal CE value for each target will decrease the time required for optimizing the tune parameters and tries to find this CE which is optimal for a broad range of peptides to be measured in the SRM assay [102].

5.5 Charge prediction

With electrospray ionization (ESI) – based LC-MS/MS, the same peptide can be ionized with different charge states. As precursor ions with different charge states have different ion intensities, an average charge state can be calculated. Prediction of these charge states is possible as the average charge state is generally influenced by the number of basic amino acid

residues in the peptide sequence. Computational models based on peptide sequence and multivariate linear regression demonstrate the ability to predict the peptide charge state in different datasets [103]. Other software, like Basophile, are based on analyzing the basicity of the N- and C-terminal fragments surrounding a peptide bond in order to predict proton segregation [98].

5.6 Optimization based on prior knowledge

In the optimal design of a proteomics experiment theoretical and statistical considerations play a fundamental role. On the other hand, in many scenarios an experimental design strategy can also benefit from a prior analysis of existing knowledge regarding the process or proteome of interest. For this task the vast amounts of information deposited in various databases and scattered in the literature can be explored, retrieved, filtered and analyzed with computational tools in the course towards setting up an experiment [104]. In the next section we will provide some hints towards which existing information can be relevant and how it could be utilized. In many cases this task can be summarized as retrieving a set of proteins that are likely to be observed with relevance to the given research question. This set can then be analyzed *in silico* using a variety of computational tools to determine their physicochemical properties in order to design a suitable experimental workflow. This allows reducing the "expected proteome" from the full theoretical proteome to a much more likely and realistic proteome. It should however be mentioned that this step comes at the risk of the so-called "bandwagon effect", if experimental design parameters are (over-) optimized to reproduce earlier observations.

5.6.1 Using previously acquired experimental proteomics datasets

Any design of proteomics experiments should be preceded by an analysis of similar studies that have previously been carried out [105]. Obviously a literature survey is not only essential to properly frame the research question in relation to the existing knowledge, but it is also essential to explore the technical possibilities for the research question (unfortunately, the literature is more efficient in telling you what could work, than in revealing what may not work). However, the low scalability and the dependence of expert interventions limits the systematic use of literature information for proteomics experiment design. Below we will discuss resources that can help to answer two important questions regarding the study that needs to be designed: 1) what do we already know; and 2) what do we expect to observe. Answers to these questions are

indispensable to maximize the amount of relevant, trustworthy and new information from a new experiment.

Organized data resources that can be relevant for systematical experiment design are the public proteomics databases and repositories [106]. For shotgun proteomics the most important public data repositories include, among others, the PRoteomics IDEntifications database (PRIDE) [107] and the Mass spectrometry Interactive Virtual Environment (MassIVE), both in the context of the ProteomeXchange consortium [108], the Global Proteome Machine Database (GPMDB) [109] , and PeptideAtlas [110]. These databases contain experimental spectra and protein identifications for a vast range of model and other organisms. Additionally, for SRM data there is the PeptideAtlas SRM Experiment Library (PASSEL) [111], also in the context of the ProteomeXchange consortium. More specific, curated resources for the extraction of organism and organelle- or biofluid-specific proteomes are for example the Human Proteinpedia [112], MAPU [113] and the Yeast Resource Center Public Data Repository [114]. For a comparative review of the major data repositories and their features we refer to several reviews [115,116,106].

Another source of public data comes in the form of spectral libraries [117]. Spectral libraries are generally used as an alternative approach to sequence database searching to identify fragment spectra. However, because spectral libraries explicitly consist of representative spectra for peptides that have been confidently identified in past experiments, they effectively form a concise representation of the proteome. Furthermore, depending on the metadata retained in the spectral libraries, the representative spectra can be linked back to their individual experiments with their respective experimental set-up and conditions.

These resources can be of great value to perform a meta-analysis of the known status of a given proteome and can answer the question which proteins can be expected to be observed under given experimental circumstances or in a specific biological context. Their usefulness for experimental design is largely dependent on the quality and accessibility of their experimental metadata: only with sufficient and well-organized metadata can all relevant existing experimental data efficiently be retrieved and interpreted from a repository. Alternatively, these data sources are also at the basis of many machine learning based models [86] that predict various properties of a theoretical proteome. In this way they are indirectly used in many of the predictive approaches covered earlier in this review.

5.6.2 Using additional knowledge of the expected proteome

A second level in which prior knowledge can be incorporated in the design of proteomics experiments, is by using annotation databases. These databases collect experimentally proven or otherwise inferred links between a given protein and specific terms that imply a function, localization or other property of the protein. Usually the terms are part of a controlled vocabulary, i.e. they have a well-defined meaning that is curated by experts. Annotation databases can be used to search for protein entries that are associated with a specific experimental context. In the context of experiment design, it could for example be relevant to extract all proteins that are linked to a given biological compartment of interest. The most obvious resource for annotation analysis is the Gene Ontology Consortium [118] (GO), but pathway databases such as KEGG [119] and reactome [120] can also be used to extract functionally relevant proteomes for an in silico analysis prior to experiment design. A powerful tool that provides systematic access to several of the aforementioned resources, and many others, is BioMart [121]. It allows performing queries over many resources through a unified interface, both in interactive (through a user-friendly interface) and programmatic ways. Together these resources allow to extract identifiers of proteins that could be expected. These identifiers can then be used to extract sequences and other features from resources such as Uniprot (The UniProt Consortium, 2011) for subsequent analysis.

An important caveat is that although annotated genomes (and thus a set of coding sequences) of most model organisms used in biomedical research are available in public databases, identification of proteins from organisms with non-sequenced genomes still remains challenging. Indeed, as protein identification requires the matching of tandem mass spectra of (usually) tryptic peptides to genomes held in (public) databases, unlocking the sequence identity of organisms that are not sequenced yet requires a different approach. Although this has no consequences for the experimental design itself, it does influence the data processing. Therefore, this should be firmly kept in mind, as computational optimization methods might not take this into account directly.

2. expert commentary

Proteomics experiments are not that different from other high-throughput omics studies when it comes to the statistical design of the experiments. Yet the optimization of the experiments proper requires specific knowledge and expertise, which in turn impacts the study design. As a result, the generalized statistical considerations that are applicable to any omics study need to be interpreted in the context of proteomics, along with the concomitant issues. The choice between a DDA, DIA or targeted proteomics approach for instance, is very fundamental to a study, but is influenced much more directly by the strengths and weaknesses of these methods than by the statistical considerations that each type of analysis requires. Indeed, it can be argued that proteomics study design is typically determined by the technological or biological limitations rather than by the statistics. As a result, it is very important for proteomics researchers to be aware of a wide variety of possible study designs, and to have access to statisticians with sufficient proteomics-specific domain knowledge.

Moreover, because of the enormous diversity in physico-chemical properties and abundances that are represented in the proteome, it can be enormously beneficial to build on previous knowledge. This can take the form of exploiting previously gleaned optimal experimental conditions, either straight from the literature, or by comparing the properties of similar data sets that have been deposited in the public domain. The most promising re-use of publicly available proteomics data for this purpose however, lies in the successful abstraction of the knowledge in predictive models. Such models are particularly interesting because these can learn from existing data to predict the properties and behavior of as-yet unseen analytes. As such, these predictive models allow novel sample types, novel analytes, and novel study designs to be first tested *in silico*.

As an overall point of attention however, it was recently reported that very many studies in the life sciences today perform very poorly when critically assessed for bias [122]. While not a problem of proteomics *per se*, it is telling that the level of statistical rigor in the life sciences leaves something to be desired. For the field of proteomics, which is after all an analytical field in essence, it is therefore important to focus much more intensely on designing adequately powered, well-considered experiments. However, a powerful force that seems to be acting against this increased level of rigor is the ability to publish findings across the impact factor space regardless of correct design. Clearly, there is a role set aside for the community at large,

to mandate more rigor, and to educate both authors and reviewers sufficiently to both design good experiments, and to detect poorly designed ones.

3. 5 years perspective

The proper design of proteomics experiments has increasingly become a focus point in the community. This shift in focus from technology development to consolidation of the technique is a signal of technology maturation. This same maturation is evident in the increased importance of quality control and quality assurance, which are the analytical siblings of experimental design [123,107]. Yet in order to successfully adopt more rigorous experimental practices, it is clear that researchers in the field need to be better educated with respect to statistics.

Indeed, education really is the elephant in the room, and should therefore take center stage. Researchers tend not to utilize suboptimal experimental designs out of malice, but rather out of ignorance. It is highly illustrative that efforts to educate researchers at the postgraduate level, such as review articles or dedicated tutorial-style articles (e.g., by the Nature Publishing Group; http://www.nature.com/collections/qghhqm) typically cover very elementary concepts. This indicates a fundamental lack of understanding that stems from the undergraduate training of researchers in the life sciences. Nevertheless, most, if not all, curricula that lead to life sciences degrees will implement one or more statistics courses, as well as analytics courses in which the statistics should be applied. Yet despite the time and importance dedicated to these courses in the vast majority of curricula, the tutorial examples cited above clearly demonstrate that the practical working knowledge of researchers at the graduate level is lacking. Perhaps this is due to the way in which statistics are taught, with a focus on the theory rather than the application, and certainly with too limited repetition of the material throughout the remainder of the curriculum. Put in another way, how many students design their own experiments in their lab work assignments?

This is perhaps the most relevant take-home message: the education of life scientists, and especially of those that will depend on complex, high-throughput analytics, should deliver on three key points: (i) to instill into these students the importance of correct experimental design from the very start; (ii) to provision the students with basic knowledge and fit-for-purpose tools to allow the design of good experiments; and (iii) to hone this training into root skills by repeated practice throughout their education.

Correct experimental design is of crucial importance for the applicability and longevity of results from the life sciences. The field of proteomics, as one of the high-throughput, molecular analytics disciplines, is directly confronted by this fundamental requirement. It is not however, sufficient to acknowledge the importance of experimental design; we also need to ensure that future proteomics researchers will be fully equipped to tackle this essential challenge. And for this, we foremost need to consider how we train these researchers.

Reference List

1. Fisher RA. The design of experiments. 1974.

2. Petricoin III EF, Ardekani AM, Hitt BA et al. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet. 2002;9306:572-577.

3. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics. 2004;5:777-785.

4. Sorace JM,Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. 2003;1.

5. Baggerly KA, Coombes KR, Morris JS. Bias, randomization, and ovarian proteomic data: a reply to "producers and consumers". Cancer Inform. 2005;9-14.

6. Hu J, Coombes KR, Morris JS et al. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. Brief.Funct.Genomic.Proteomic. 2005;4:322-331.

7. Bittremieux W, Willems H, Kelchtermans P et al. iMonDB: Mass Spectrometry Quality Control through Instrument Monitoring. J Proteome Res. 2015;5:2360-2366.

8. Nesvizhskii AI,Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics. 2005;10:1419-1440.

9. He Z, Huang T, Liu X et al. Protein inference: A protein quantification perspective. Computational Biology and Chemistry.

10. Thompson AJ, Abu M, Hanger DP. Key issues in the acquisition and analysis of qualitative and quantitative mass spectrometry data for peptide-centric proteomic experiments. Amino Acids. 2012;3:1075-1085.

11. Cappadona S, Baker PR, Cutillas PR et al. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Amino Acids. 2012;3:1087-1108.

12. Lemeer S, Hahne H, Pachl F et al. Software tools for MS-based quantitative proteomics: a brief overview. Methods Mol Biol. 2012;489-499.

13. Serang O,Kall L. Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. J Proteome Res. 2015;10:4099-4103.

14. Mason RL, Gunst RF, Hess JL. Nested Designs. 2003;378-399.

15. Mallick P,Kuster B. Proteomics: a pragmatic perspective. Nat Biotech. 2010;7:695-709.

16. Canas B, Pieiro C, Calvo E et al. Trends in sample preparation for classical and second generation proteomics. Journal of Chromatography A. 2007;1ΓÇô2:235-258.

17. Mottaz-Brewer HM, Norbeck AD, Adkins JN et al. Optimization of Proteomic Sample Preparation Procedures for Comprehensive Protein Characterization of Pathogenic Systems. J Biomol Tech. 2008;5:285-295.

18. Wisniewski JR, Zougman A, Mann M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J Proteome Res. 2009;12:5674-5678.

19. Glibert P, Van SK, Dhaenens M et al. iTRAQ as a method for optimization: enhancing peptide recovery after gel fractionation. Proteomics. 2014;6:680-684.

20. Vandermarliere E, Mueller M, Martens L. Getting intimate with trypsin, the leading protease in proteomics. Mass Spec Rev. 2013;6:453-465.

21. Meyer JG, Kim S, Maltby DA et al. Expanding proteome coverage with orthogonal-specificity alpha-lytic proteases. Mol Cell Proteomics. 2014;3:823-835.

22. Saveliev S, Bratz M, Zubarev R et al. Trypsin/Lys-C protease mix for enhanced protein mass spectrometry analysis. Nat Meth. 2013;11.

23. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res. 2010;3:1323-1329.

24. Thakur SS, Geiger T, Chatterjee B et al. Deep and Highly Sensitive Proteome Coverage by LC-MS/MS Without Prefractionation. Mol Cell Proteomics. 2011;8.

25. Nagaraj N, Kulak NA, Cox J et al. System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. Mol Cell Proteomics. 2012;3.

26. Ahmed FE. Liquid chromatography-mass spectrometry: a tool for proteome analysis and biomarker discovery and validation. Expert Opin.Med.Diagn. 2009;4:429-444.

27. Mostovenko E, Hassan C, Rattke J et al. Comparison of peptide and protein fractionation methods in proteomics. EuPA Open Proteomics. 2013;30-37.

28. Meysman P, Titeca K, Eyckerman S et al. Protein complex analysis: From raw protein lists to protein interaction networks. Mass Spectrom.Rev. 2015.

29. Millioni R, Tolin S, Puricelli L et al. High abundance proteins depletion vs low abundance proteins enrichment: comparison of methods to reduce the plasma proteome complexity. PLoS One. 2011;5:e19603.

30. Smolders K, Lombaert N, Valkenborg D et al. An effective plasma membrane proteomics approach for small tissue samples. Sci.Rep. 2015;10917.

31. Klie S, Martens L, Vizcaino JA et al. Analyzing large-scale proteomics projects with latent semantic indexing. J Proteome Res. 2008;1:182-191.

32. Gallien S, Duriez E, Demeure K et al. Selectivity of LC-MS/MS analysis: implication for proteomics experiments. J Proteomics. 2013;148-158.

33. Gallien S,Domon B. Advances in high-resolution quantitative proteomics: implications for clinical applications. Expert Rev Proteomics. 2015;5:489-498.

34. Kim YJ, Gallien S, van OJ et al. Targeted proteomics strategy applied to biomarker evaluation. Proteomics Clin.Appl. 2013;11-12:739-747.

35. Domon B,Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol. 2010;7:710-721.

36. Wisniewski JR, Dus K, Mann M. Proteomic workflow for analysis of archival formalin-fixed and paraffin-embedded clinical samples to a depth of 10 000 proteins. Proteomics Clin.Appl. 2013;3-4:225-233.

37. Egertson JD, Kuehn A, Merrihew GE et al. Multiplexed MS/MS for improved data-independent acquisition. Nat Methods. 2013;8:744-746.

38. Egertson JD, Kuehn A, Merrihew GE et al. Multiplexed MS/MS for improved data-independent acquisition. Nat Methods. 2013;8:744-746.

39. Rost HL, Rosenberger G, Navarro P et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 2014;3:219-223.

40. Gillet LC, Navarro P, Tate S et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012;6:O111.

41. Vaudel M, Sickmann A, Martens L. Peptide and protein quantification: a map of the minefield. Proteomics. 2010;4:650-670.

42. Thompson A, Schafer J, Kuhn K et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal.Chem. 2003;8:1895-1904.

43. Chong PK, Gan CS, Pham TK et al. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections. J Proteome Res. 2006;5:1232-1240.

44. Ong SE,Mann M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat.Protoc. 2006;6:2650-2660.

45. Gerber SA, Rush J, Stemman O et al. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc.Natl.Acad.Sci.U.S.A. 2003;12:6940-6945.

46. Brownridge PJ, Harman VM, Simpson DM et al. Absolute multiplexed protein quantification using QconCAT technology. Methods Mol Biol. 2012;267-293.

47. Zhang Y, Fonslow BR, Shan B et al. Protein analysis by shotgun/bottom-up proteomics. Chem.Rev. 2013;4:2343-2394.

48. Vaudel M, Burkhart JM, Radau S et al. Integral quantification accuracy estimation for reporter ion-based quantitative proteomics (iQuARI). J Proteome Res. 2012;10:5072-5080.

49. Christoforou A,Lilley KS. Taming the isobaric tagging elephant in the room in quantitative proteomics. Nat Methods. 2011;11:911-913.

50. Nahnsen S,Kohlbacher O. In silico design of targeted SRM-based experiments. 2012;Suppl 16.

51. Picotti P,Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Meth. 2012;6:555-566.

52. Zubarev R,Mann M. On the proper use of mass accuracy in proteomics. Mol Cell Proteomics. 2007;3:377-381.

53. Zhang Y, Wen Z, Washburn MP et al. Effect of dynamic exclusion duration on spectral count based quantitative proteomics. Anal Chem. 2009;15:6317-6326.

54. Vissers JP, Blackburn RK, Moseley MA. A novel interface for variable flow nanoscale LC/MS/MS for improved proteome coverage. J Am.Soc.Mass Spectrom. 2002;7:760-771.

55. Batth TS, Singh P, Ramakrishnan VR et al. A targeted proteomics toolkit for high-throughput absolute quantification of Escherichia coli proteins. Metab Eng. 2014;48-56.

56. Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics. 2012;S5.

57. Lai X, Wang L, Witzmann FA. Issues and applications in label-free quantitative mass spectrometry. Int.J Proteomics. 2013;756039.

58. Han J, Ye L, Xu L et al. Towards high peak capacity separations in normal pressure nanoflow liquid chromatography using meter long packed capillary columns. Anal Chim.Acta. 2014;267-273.

59. Aebersold R,Mann M. Mass spectrometry-based proteomics. Nature. 2003;6928:198-207.

60. Mee R. A comprehensive guide to factorial two-level experimentation. Springer Science & Business Media, 2009.

61. Dean A, Lewis, S. Screening: methods for experimentation in industry, drug discovery, and genetics. Springer Science & Business Media. 2006.

62. Oberg AL,Vitek O. Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments. J.Proteome Res. 2009;5:2144-2156.

63. Karp NA,Lilley KS. Design and Analysis Issues in Quantitative Proteomics Studies. Proteomics. 2007;S1:42-50.

64. Vaudel M, Barsnes H, Martens L et al. Bioinformatics for Proteomics: Opportunities at the Interface Between the Scientists, Their Experiments, and the Community. 2014;239-248.

65. Morris J, Baggerly KA, Gutstein HB et al. Statistical Contributions to Proteomic Research. 2010;143-166.

66. Kendziorski C, Irizarry RA, Chen KS et al. On the utility of pooling biological samples in microarray experiments. Proc.Natl.Acad.Sci.U.S.A. 2005;12:4252-4257.

67. Peng X, Wood CL, Blalock EM et al. Statistical implications of pooling RNA samples for microarray experiments. BMC Bioinformatics. 2003;26.

68. Diz AP, Truebano M, Skibinski DO. The consequences of sample pooling in proteomics: an empirical study. Electrophoresis. 2009;17:2967-2975.

69. Karp NA, Spencer M, Lindsay H et al. Impact of replicate types on proteomic expression analysis. J Proteome Res. 2005;5:1867-1871.

70. Raji MA, Schug KA. Chemometric study of the influence of instrumental parameters on ESI-MS analyte response using full factorial design. Int. J. Mass Spectrom. 2009; 279(2/3): 100–106.

71. Andrews GL, Dean RA, Hawkridge AM et al. Improving proteome coverage on a LTQ-Orbitrap using design of experiments. J Am.Soc.Mass Spectrom. 2011;4:773-783.

72. Prieto A, Zuloaga O, Usobiaga A et al. Development of a stir bar sorptive extraction and thermal desorption-gas chromatography-mass spectrometry method for the simultaneous determination of several persistent organic pollutants in water samples. J Chromatogr.A. 2007;1-2:40-49.

73. Sutton RS. Reinforcement learning: an introduction. 1998;322.

74. Auer P, Cesa-Bianchi N, Fischer P. Finite-time Analysis of the Multiarmed Bandit Problem. 2002;2-3:235-256.

75. Settles B. Active Learning. 2012;1:1-114.

76. Jones DR, Schonlau M, Welch WJ. Efficient Global Optimization of Expensive Black-Box Functions. 1998;4:455-492.

77. Jones DR. A Taxonomy of Global Optimization Methods Based on Response Surfaces. 2001;4:345-383.

78. Santner TJ. The Design and analysis of computer experiments. 2003;283.

79. Knowles J. ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. 2006;1:50-66.

80. Gorissen D, Couckuyt I, Demeester P et al. A Surrogate Modeling and Adaptive Sampling Toolbox for Computer Based Design. 2010;2051-2055.

81. Forrester AIJ,Keane AJ. Recent advances in surrogate-based optimization. 2009;1-3:50-79.

82. Verbeeck D, Maes F, De Grave K et al. Multi-objective optimization with surrogate trees. 2013.

83. De Grave K, Ramon J, De Raedt L. Active Learning for High Throughput Screening. 2008;185-196.

84. Dunham WH, Larsen B, Tate S et al. A costΓÇôbenefit analysis of multidimensional fractionation of affinity purification-mass spectrometry samples. Proteomics. 2011;13:2603-2612.

85. Holstein Sherwood CA, Gafken PR, Martin DB. Collision energy optimization of b- and y-ions for multiple reaction monitoring mass spectrometry. J Proteome Res. 2011;1:231-240.

86. Kelchtermans P, Bittremieux W, De Grave K et al. Machine learning applications in proteomics research: How the past can boost the future. Proteomics. 2014;4-5:353-366.

87. Sun Y, Braga-Neto U, Dougherty ER. A systematic model of the LC-MS proteomics pipeline. BMC Genomics. 2012;Suppl 6.

88. Fannes T, Vandermarliere E, Schietgat L et al. Predicting Tryptic Cleavage from Proteomics Data Using Decision Tree Ensembles. J.Proteome Res. 2013;5:2253-2259.

89. Baczek T,Kaliszan R. Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. Proteomics. 2009;4:835-847.

90. Meek JL. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. Proc.Natl.Acad.Sci.U.S.A. 1980;3:1632-1636.

91. Mant CT,Hodges RS. Context-dependent effects on the hydrophilicity/hydrophobicity of side-chains during reversed-phase high-performance liquid chromatography: Implications for prediction of peptide retention behaviour. J Chromatogr.A. 2006;2:211-219.

92. Klammer A, Yi X, MacCoss MJ, Noble WS. Peptide Retention Time Prediction Yields Improved Tandem Mass Spectrum Identification for Diverse Chromatography Conditions. Research in Computational Molecular Biology. 2009;4453:459-472.

93. Spicer V, Yamchuk A, Cortens J et al. Sequence-specific retention calculator. A family of peptide retention time prediction algorithms in reversed-phase HPLC: applicability to various chromatographic conditions and columns. Anal Chem. 2007;22:8762-8768.

94. Moruz L, Tomazela D, Ka¦êll L. Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. J.Proteome Res. 2010;10:5209-5216.

95. Moruz L, Staes A, Foster JM et al. Chromatographic retention time prediction for posttranslationally modified peptides. Proteomics. 2012;8:1151-1159.

96. Parker SJ, Rost H, Rosenberger G et al. Identification of a Set of Conserved Eukaryotic Internal Retention Time Standards for Data-independent Acquisition Mass Spectrometry. Mol Cell Proteomics. 2015;10:2800-2813.

97. Li YF, Arnold RJ, Tang H et al. The Importance of Peptide Detectability for Protein Identification, Quantification, and Experiment Design in MS/MS Proteomics. J.Proteome Res. 2010;12:6288-6297.

98. Wang D, Dasari S, Chambers MC et al. Basophile: accurate fragment charge state prediction improves peptide identification rates. Genomics Proteomics Bioinformatics. 2013;2:86-95.

99. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. Anal Chem. 2005;19:6364-6373.

100. Arnold RJ, Jayasankar N, Aggarwal D et al. A machine learning approach to predicting peptide fragmentation spectra. Pacific Symposium on Biocomputing. 2006;11.

101. Degroeve S, Maddelein D, Martens L. MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. Nucleic Acids Res. 2015;W1:W326-W330.

102. MacLean B, Tomazela DM, Abbatiello SE et al. Effect of Collision Energy Optimization on the Measurement of Peptides by Selected Reaction Monitoring (SRM) Mass Spectrometry. Anal.Chem. 2010;24:10116-10124.

103. Liu H, Zhang J, Sun H, Xu C. The prediction of peptide charge states for electrospray ionization in mass spectrometry. Procedia Environmental Sciences 2011;8:483-491.

104. Barsnes H,Martens L. Crowdsourcing in proteomics: public resources lead to better experiments. Amino Acids. 2013;4:1129-1137.

105. Vaudel M, Verheggen K, Csordas A et al. Exploring the potential of public proteomics data. Proteomics. 2016;2:214-225.

106. Perez-Riverol Y, Wang R, Hermjakob H et al. Open source libraries and frameworks for mass spectrometry based proteomics: A developer's perspective. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2014;1, Part A:63-76.

107. Martens L, Hermjakob H, Jones P et al. PRIDE: The proteomics identifications database. Proteomics. 2005;13:3537-3545.

108. Vizcano JA, Csordas A et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucl.Acids Res. 2013;D1:D1063-D1069.

109. Craig R, Cortens JP, Beavis RC. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. J.Proteome Res. 2004;6:1234-1242.

110. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. 2008;5:429-434.

111. Farrah T, Deutsch EW, Kreisberg R et al. PASSEL: The PeptideAtlas SRMexperiment library. Proteomics. 2012;8:1170-1175.

112. Mathivanan S, Ahmed M, Ahn NG et al. Human Proteinpedia enables sharing of human protein data. Nat Biotech. 2008;2:164-167.

113. Gnad F, Oroshi M, Birney E et al. MAPU 2.0: high-accuracy proteomes mapped to genomes. Nucl.Acids Res. 2009;suppl 1:D902-D906.

114. Riffle M, Malmstram L, Davis TN. The Yeast Resource Center Public Data Repository. Nucl.Acids Res. 2005;suppl 1:D378-D382.

115. VizcanoJA, Mueller M, Hermjakob H et al. Charting online OMICS resources: A navigational chart for clinical researchers. Prot.Clin.Appl. 2009;1:18-29.

116. Vizcano JA, Foster JM, Martens L. Proteomics data repositories: Providing a safe haven for your data and acting as a springboard for further research. Journal of Proteomics. 2010;11:2136-2146.

117. Lam H,Aebersold R. Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. Methods. 2011;4:424-431.

118. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.Genet. 2000;1:25-29.

119. Kanehisa M,Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;1:27-30.

120. Matthews L, Gopinath G, Gillespie M et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009;Database issue:D619-D622.

121. Smedley D, Haider S, Ballester B et al. BioMart--biological queries made easy. BMC Genomics. 2009.

122. Macleod MR, Lawson MA, Kyriakopoulou A et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLoS Biol. 2015;10:e1002273.

123. Tabb DL. Quality assessment for clinical proteomics. Clin.Biochem. 2013;6:411-420.