Bioanalytical Omics

Subgroup report

Renate König, DE (subgroup lead)
Alison Cave, EMA
Mark Goldammer, DE
Didier Meulendijks, NL

See websites for contact details
Heads of Medicines Agencies www.hma.eu
European Medicines Agency www.ema.europa.eu

An agency of the
European Union

Table of content

1. Summary

The section focuses on the characterisation and mapping of bioanalytical omics data, specifically proteomics, metabolomics and lipidomics data. Traditionally, Big Data/Omics approaches are used for hypothesis generation and thus, used in basic research – with minor impact on regulatory issues. However, there is significant development and potential that these technologies will be applied also with (hypothesis driven) confirmatory purpose for the development and application of medicines. Consequently, there is the need to analyse and follow-up on the ongoing development in order to prepare the regulatory systems for the potential impact of Bioanalytical omics data. In principal, proteomics is a technique that holds great promise and can offer approaches supporting the development of infrastructures for personalised medicine. Bioanalytical Omics technologies, besides proteomics, such as metabolomics and lipidomics, could potentially deliver a dynamic stratification of patient populations over time rather than a single snapshot measurement. Data from these fields are highly variable and while current initiatives have been formed to address these issues there is a regulatory need to sufficiently ascertain the validity of bioanalytical omics data. Validation of bioanalytical methods for proteomics, lipidomics and metabolomics is highly challenging, when compared to validation of other types of bioanalytical methods. There is a need to define the scope and extent of validation necessary for these methods, and how the standards should be set in order to ensure appropriate data quality for use in regulatory processes. Additionally, there is large heterogeneity among data formats and data analysis pipelines in the fields of proteomics, metabolomics and lipidomics. From a regulatory point of view, it is also important to define the requirements or standards to which databases/data standards should adhere. Moreover, bioinformatic analysis of proteomics data requires highly specific methodology. Current approaches combining genomics and proteomics do already rely on bioinformatics algorithms for target prediction as opposed to direct identification by proteomics. For example, active personalisation in therapeutic cancer vaccination relies on proteomics analysis (see key case study in this report). In case these algorithms will inform therapeutic decisions, or their use is intended for the evaluation of efficacy and safety of medicines, their robustness and predictive value has to be ensured. From a regulatory perspective, there is a need to define how to assess and validate algorithms and the applied statistical methods and/or approaches. Thus, regulatory qualification of these methods is encouraged to (i) improve the quality and comparability of results; and (ii) improve the preparedness of regulatory systems for appropriate assessment of results.

2. Background / Scope

The development of high-throughput technologies with the potential to comprehensively assess not only the genome but also the proteome and other molecular markers provide the possibility to identify unique molecular markers not only of disease but also for responsiveness to medications. This report focuses on the characterisation of "bioanalytical omics" including Proteomics, Metabolomics and Lipidomics, three fields of research/technologies, that are largely mass spectrometry-driven, whereas Genomics technologies (driven by sequencing) will be addressed in a different section of the report on Big Data. Where possible it will provide concrete examples of technologies that already have contributed to regulatory processes, to date mostly in the field of proteomics, and will provide an outlook on future technologies that are expected to have a substantial impact on regulatory processes in the near future.

2.1. Definition of Big Data

Although the phenomenon of Big Data is widely acknowledged there is not one commonly accepted definition of Big Data. We refer to Big Data as are data that contain relevant information to answer a

given question but that will in general require advanced or specialized methods to provide a suitable answer within reliable constraints.

Such methods of data analysis (i.e. models and algorithms) are subject to the growing field of data science which combines methods from various disciplines such as biostatistics, mathematical modelling and simulation, bioinformatics and computer science including data-integration, machine learning and high-performance computing. Data and analytic methods that will be relevant in the regulatory context will be exemplarily discussed within this document. The use of Big Data and their analysis call for new regulatory strategies and guidance to achieve their full beneficial potential.

2.2. Proteomics

Proteomics is the comprehensive study of a specific proteome resulting from its genome, including abundances and quantifications of proteins, their variations and modifications, interacting partners and networks. The proteome is the entire set of expressed proteins in a given type of cell, tissue or organism, at a given time, under defined conditions. Much focus over the past 20 years has been on using sequencing to assess genome and transcriptome information. However, genomics and transcriptional profiles are not always reliable predictors of protein levels or activity. Hence, given that gene products like proteins – and not genes – are the main targets of most therapeutics, quantifying proteins in health and disease is of significant importance. However, the proteome is complex; many RNA transcripts result in more than one protein due to alternative splicing, allelic variations or alternative post-translational modifications (PTM). Complicating the situation is that the latter have a wide range of effects, broadening the range of functionality of the protein (Khoury et al. 2011). Whereas mutations can only occur once per position, different forms of PTMs may occur in tandem, are dynamic, and currently there is no method to readily assess their relative levels. Currently, the rate of detection of PTM is far outstripping our ability to assign biological function to these modifications. For comparison, the human genome contains about 20,300 protein-encoding genes, but the total number of proteins in human cells are proposed to outreach several millions (Ponomarenko et al. 2016) up to estimates of even billions of proteoforms (Smith und Kelleher 2013).

There are two main types of proteomic studies: discovery proteomics and targeted proteomics which can be performed on a range of sample types including specific clinical biospecimens or samples from pre-clinical models such as animal models or cultured cells. Discovery proteomics refers to "untargeted" identification and quantification of proteins in a biological or clinical sample while quantitative measurements on a defined subset of total proteins in a biological or clinical sample are termed targeted proteomics. Commonly used high-throughput technologies encompass different types of mass spectrometry (MS), protein chips and reverse-phase protein microarrays. The latter can be specifically used in combination with laser capture microdissection to capture various stages of diseases within an individual patient or compare diseased and healthy tissues within the same patient to develop treatment strategies. Complex bioinformatic tools are an integral and essential component of proteomic studies.

Proteogenomics research (in which proteomics and genomics data are combined) is a relatively new "omics" field, with a significant potential to guide future approaches, and needs to be discussed. By combining genomic (e.g. NGS) and proteomic (e.g. mass spectrometry) information, the information gain is much higher, and this is expected to have a substantial impact on future clinical "omics-based" research, in oncology for example in defining genomic/proteomic signatures of human tumours (see case study below in chapter 6)

In order to understand the genome, a firm understanding of the proteome is needed. This is not trivial, but there are multiple ongoing efforts to map large scale protein-protein interaction data and establish comprehensive databases which will aid the interpretation of the vast genomic datasets (Li

und Liao 2017). However, the analysis of the resulting multidimensional data sets and complex scientific problems require novel algorithm-based methodologies. Exploring the biological relevance of detected associations and subsequently validating those findings and establishing robust and predictive methods is even more challenging. As outlined below, there are a number of bioanalytical and biometric challenges that have to be addressed for establishing the acceptability of proteomics-based analytical methods in the regulatory context. However, this approach has the potential to improve medicines development processes and to significantly improve the availability of innovative and effective treatments in the future.

2.3. Metabolomics

The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the (end) products of cellular processes, including for instance circulating small molecules, such as amino acids, lipids (fats), nucleotides, and carbohydrates. These products of cellular processes provide an immediate barometer of cellular physiology. However, the metabolome is also extremely complex with a vast number of metabolites varying dramatically in chemical structure and concentration, but also changing rapidly over time e.g. due to degradation processes. The challenge in metabolomics is to find a bioanalytical approach, which would allow the reproducible identification of specific metabolites and quantitation of metabolites across multiple different contexts. The data complexity is particularly problematic for metabolomics, which is characterised by more heterogeneity resulting in a more complex search space than bioanalytical omics technologies due to the much wider range of molecular entities measured.

Metabolomics is the scientific study of chemical processes involving metabolites, the unique "fingerprint" that cellular processes leave behind. A related term is "pharmaco-metabolomics", which is defined as "the prediction of the effects of a drug on the basis of a mathematical model of pre-dose metabolite profiles" (Everett et al. 2013). The key advantage of metabolomics is that whilst the genomic DNA sequence of an individual is fixed, the metabotype (metabolic phenotype) offers the possibility of capturing over time the influence of environmental factors and other factors on an individual's response to a medicine. Given that it is estimated that social, behavioural and environmental factors account for 70% of the factors that influence our health, technologies that offer the promise of capturing this information could be relevant for regulatory decision making. Bioanalytical technologies for metabolomics include NMR-spectroscopy and MS-based approaches. Specifically, methods that do not require the application of a matrix in order to facilitate small-molecule identification, for instance secondary ion mass spectrometry (SIMS), desorption electrospray ionisation (DESI) or laser ablation ESI (LAESI).

Additionally, imaging approaches such as PET (Positron emission tomography) allow analysing the physiological locations of binding sites of ligands and analysis of bio-distribution in the living body, including brain. These imaging approaches will be addressed in a different section of the report on Big Data and are out of the scope for this section on bioanalytical omics but should be considered as another 'bioanalytical omics' approach.

2.4. Lipidomics

Lipidomics research involves the identification and quantification of the complete lipid profile within a cell or tissue (the lipidome), which may represent thousands of cellular metabolites and their interactions with other lipids, proteins, and other metabolites. While the lipidome represents a subset of the metabolome, which also includes the three other major classes of biological molecules (proteins/amino-acids, sugars and nucleic acids), it is itself generally considered a distinct discipline, due to the uniqueness and functional specificity of lipids relative to other metabolites. Lipids are

essential metabolites that, as for the larger broader metabolome, provide readout of the cellular metabolic status. They are either generated endogenously or incorporated into cells from dietary sources. In particular multiple human pathologies, including common chronic conditions partially with an underpinning inflammatory component such as cardiovascular disease, diabetes and neurodegenerative diseases involve changes in lipid metabolism. As such, there is an increasing interest in the development of medicines which directly or indirectly target the lipid metabolic pathways and in this respect the analysis of the lipidome may potentially provide a useful clinical diagnostic method. Commonly used high-throughput technologies in lipidomics such as NMR spectroscopy, fluorescence spectroscope, dual polarization polarization interferometry and mass spectrometry, such as ESI, DESI and matrix-assisted laser desorption/ionization ionization (MALDI) may open avenues to the role of specific lipidomic profiles in the development and progression of disease metabolites within (common) pathways (Reviews on lipidomics and applications: (Vihervaara et al. 2014; Triebl et al. 2017).

3. Objectives

- To map relevant sources of "bioanalytical omics" data and examples of data formats with focus on proteomics.
- To discuss issues related to data quality and variability.
- To identify requirements and potential pathways for the regulatory validation/qualification – paving the way for future applications.
- To illustrate regulatory challenges in a key case study evaluating application in personalised medicine approaches.

4. Methods

Public databases and peer-reviewed publications (as indicated) were considered in this report. Furthermore, input was received from experts at the Paul-Ehrlich-Institute, the Federal Institute for Vaccines and Biomedicines, Germany, that are working in the fields of Therapeutic Vaccines (Dr. Thomas Hinz), Monoclonal and Polyclonal Antibodies (Dr. Jörg Engelbergs), Biostatistics and Modelling (Dr. Christel Kamp; Dr. Gaby Wangorsch).

5. Data and method characterisation / qualification

5.1. General overview of methodology

5.1.1. Proteomics – Quantitative Analysis Technology

Quantification of proteins has historically been performed using immunoassays; however, these are difficult to multiplex across multiple laboratories are semi-quantitative at best (Paulovich und Whiteaker 2016). In addition, these assays are dependent on the quality of the available antibodies, which often have variable reliability, specificity and sensitivity. As a result of these limitations, mass spectrometry is now the main quantitative technique for most "omics" approaches considered within this report and therefore the focus in this section will be on this technology for performing proteomic analyses.

Tandem mass spectrometry (MS/MS) is the main driver for data dependent acquisition (DDA) as a discovery engine in proteomics. The technology allows for simultaneous sequencing of the complete proteome in a biological sample. However, reliable quantitative assays to follow up a predetermined

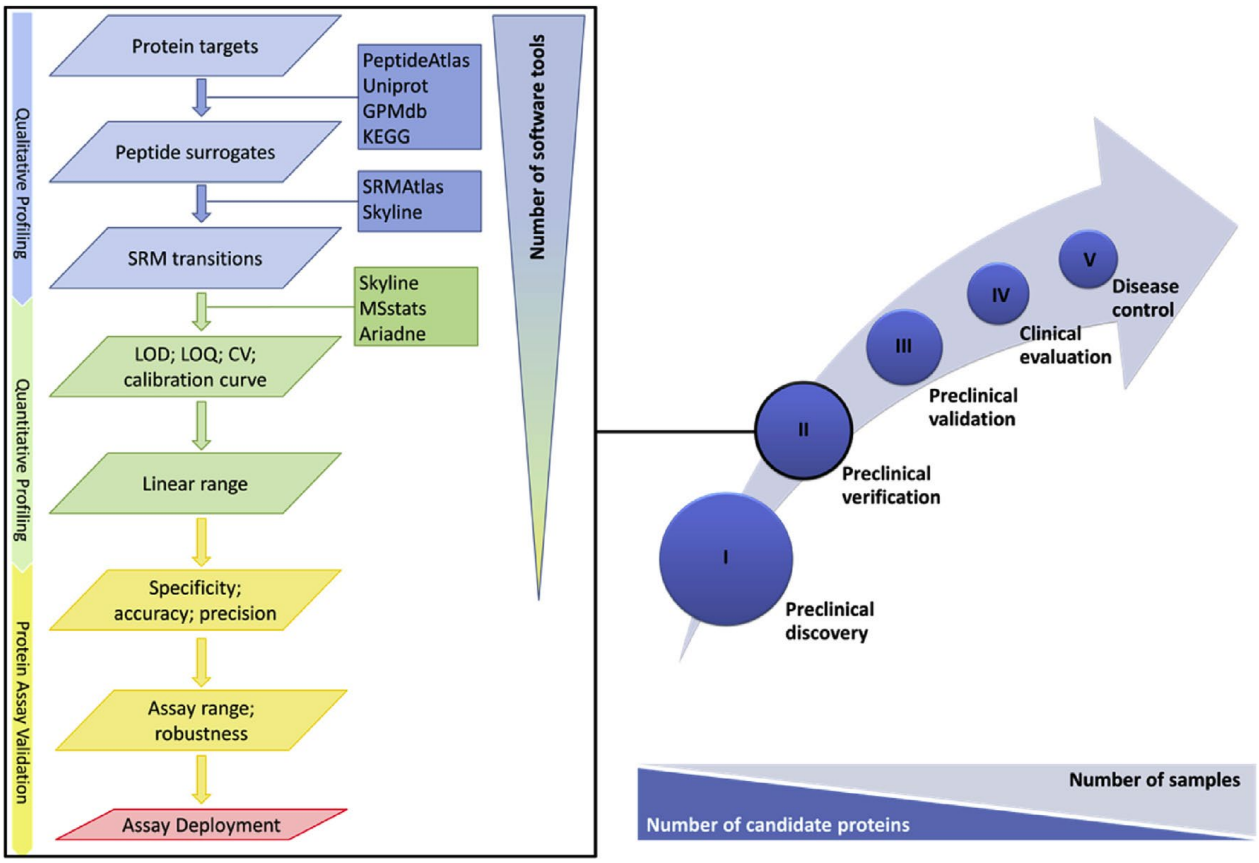
set of proteins are essential. In line with the ideal criteria of a biomarker, basic performance metrics of the bioanalytical assays used need to be established, and provide information on precision, repeatability, reproducibility, bias, linearity over a wide range of concentrations, the limits of quantification, matrix effects and selectivity and analyse stability (Grant und Hoofnagle 2014; Vidova und Spacil 2017); EMA guideline on bioanalytical method validation: EMEA/CHMP/EWP/192217/2009).

In discovery (shotgun) proteomics, a MS instrument typically operates in DDA mode, the relative abundances of peptides (products of protein enzymatic digestion) are estimated by spectral counting – the number of MS/MS spectra assigned to the same peptide/protein. Alternatively, a relative abundance can be obtained as an integrated area of peptide peak in extracted ion chromatogram (XIC) using dedicated software tools for data processing. Isotopic-labelling strategies, such as isotope-coded affinity tag (ICAT) and stable isotope labelling by amino acids in cell culture (SILAC) are used for quantitative comparison across biological samples. The quantitative comparison is critical in order to understand the clinical validity and utility of a biomarker. The development of chemical labelling by isobaric tags for relative and absolute quantification (iTRAQ) and dimethyl labelling protocols has improved quantification accuracy.

However, MS technology is not inherently quantitative due to large variance in sequence-dependent relative signal intensity. In principle, DDA-based proteomics aims at achieving unbiased and complete coverage of the proteome. However, DDA is driven by the immediate relative signal intensity of a peptide; hence, DDA results in nearly stochastic sampling, which translates into inherently limited reproducibility of both spectral counting and feature-based quantification. Several comprehensive studies on human cell lines reported approximately 50% proteome coverage in DDA mode highlighting the limited sensitivity and sequencing speed of DDA to cover complex proteomes. This is likely to create significant challenges for the identification of biomarkers since, for example, structural proteins accounting for cellular machineries are obviously much more abundant than regulatory proteins (Aebersold und Mann 2016; Vidova und Spacil 2017). However, the latter are potentially of most interest in relation to certain diseases or treatments thereof.

The targeted MS/MS technique of selected reaction monitoring (SRM, sometimes referred to as multiple reaction monitoring, MRM), aims at the reproducible, sensitive and streamlined acquisition of a subset of known peptides of interest as standard for quantitative proteomics (as an alternative to immunoassays such as ELISA or Western Blot). In contrast to the discovery (or shotgun) proteomics approach with the help of DDA, the proteins of interest are predetermined and known. With the help of the pre-existing information, peptides are selectively and recursively isolated and then fragmented over their chromatographic elution time (Aebersold und Mann 2016). The SRM technique runs on a triple-quadrupole (QqQ) mass spectrometer, which allows for flexible, quantitative and relatively routine measurement of hundreds of peptides within a shorter time frame and can be used to detect analytes of interest within complex samples. In comparison to immunoassays, the SRM technique features fast assay development and deployment and is in principle capable of distinguishing highly similar 'proteoforms' such as protein variants (isoforms), post-translationally modified proteins and genetic variants such as single nucleotide polymorphisms (SNPs) (Vidova und Spacil 2017). A quantitative SRM assay as stage II in development of a biomarker test is depicted in Fig. 1 (Vidova und Spacil 2017). In proteomics, peptides become protein surrogates and carriers of quantitative information. The process of surrogate selection and evaluation requires the following steps: i) qualitative profiling, ii) quantitative profiling and iii) assay validation, as illustrated in Fig. 1.

Fig. 1: Quantitative SRM assay as stage II in development of a biomarker test (Vidova und Spacil 2017).



Another label-free quantification (LFQ) approach used in proteomics is the data-independent acquisition (DIA) strategy relying on information from pre-existing high-quality spectral libraries. In this method, entire ranges of precursors are fragmented at the same time, followed by the acquisition of the fragments in a time-of-flight mass spectrometer, aimed at generating comprehensive fragment-ion maps for a sample. The multiplexed fragment spectra are often interpreted with the help of known fragment spectra from large spectral libraries by software such as OpenSWATH (Aebersold und Mann 2016). This approach avoids an abundance-based preselection of survey ions for fragmentations and can thereby be used for comprehensive proteome-wide quantification. The advantage is the seamless and rapid analysis, which eliminates the missing value problem of DDA (Aebersold und Mann 2016).

Conclusion

The challenge of DDA is the stochastic nature of precursor selection and low sampling efficiency due to the limited speed of mass spectrometers. Even after replication of measurements, it results in missing individual peptide identification across runs within a larger dataset. DIA can overcome some of the limitations of DDA. However, DDA encompasses a wider dynamic range of detection within a complex matrix. In short, the best method depends on the purpose of the analysis. New analytical approaches are currently addressing the "missing value" issue in DDA (Zhang et al. 2016). They claim it may not be an intrinsic problem of DDA, rather the processing of the signal information needs to be changed to a new analytical workflow that recovers missing values using a protein scoring scheme for quality control.

All methods described above represent a big data approach based on statistical associations, for systematically probing the prototype, the state of a proteome that is associated with a specific phenotype. Interestingly, proteomics data sets are often linked to genomics and clinical data. For instance, in a pioneering study, analysing the proteogenomics of breast cancer, the functional consequences of somatic mutations were elucidated and identified therapeutic targets or biomarkers (Mertins et al. 2016). In essence, this study supported the relationship between proteotypes and phenotypes through association studies between genetic loci, the resulting protein network state and clinical disease phenotype.

The table below provides a summary of the strengths and limitations of the various proteomic approaches and the data analysis pipeline (Fig. 2 and 3).

Fig. 2: Advantages and Disadvantages of various techniques used in proteomics (Chandramouli und Qian 2009)**.**

TABLE 1: Common proteomic technologies, applications, and their limitations.

Technology	Application	Strengths	Limitations
2DE	Protein separation Quantitative expression profiling	Relative quantitative PTM information	Poor separation of acidic, basis, hydrophobic and low abundant proteins
DIGE	Protein separation Quantitative expression profiling	Relative quantitative PTM information High sensitivity Reduction of intergel variability	Proteins without lysine cannot be labeled Requires special equipment for visualization and fluorophores are very expensive
ICAT	Chemical isotope labeling for quantitative proteomics	Sensitive and reproducible Detect peptides with low expression levels	Proteins without cysteine residues and acidic proteins are not detected
SILAC	Direct isotope labeling of cells Differential expression pattern	Degree of labelling is very high Quantitation is straightforward	SILAC labeling of tissue samples is not possible
iTRAQ	Isobaric tagging of peptides	Multiplex several samples Relative quantification High-throughput	Increases sample complexity Require fractionation of peptides before MS
MUDPIT	Identification of protein-protein interactions Deconvolve complex sets of proteins	High separation Large protein complexes identification	Not quantitative Difficulty in analyzing the huge data set Difficult to identify isoforms
Protein array	Quantitate specific proteins used in diagnostics (biomarkers or antibody detection) and discovery research	High-throughput Highly sensitive Low sample consumption	Limited protein production Poor expression methods Availability of the antibodies Accessing very large numbers of affinity reagents
Mass spectrometry	Primary tool for protein identification and characterization	High sensitivity and specificity High-throughput Qualitative and quantitative PTM information	No individual method to identify all proteins. Not sensitive enough to identify minor or weak spots. MALDI and ESI do not favor identification of hydrophobic peptides and basic peptides
Bioinformatics	Analysis of qualitative and quantitative proteomic data	Functional analysis, data mining, and knowledge discovery from mass spectrometric data	No integrated pipeline for processing and analysis of complex data Search engines do not yield identical results

Fig. 3: An overview of proteomic strategies and data analysis pipelines (Chandramouli und Qian 2009). Note: gel-based proteomics is listed for comprehensiveness, but almost not applied anymore.

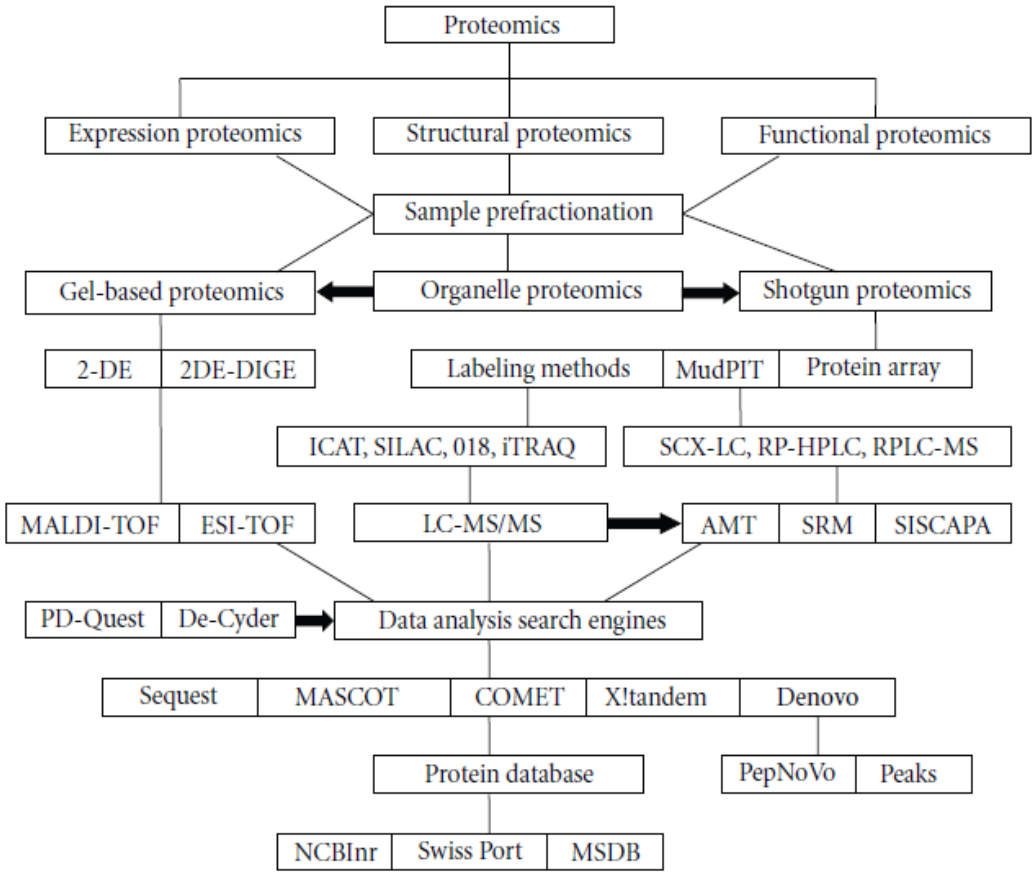


FIGURE 1: An overview of proteomic strategies.

5.1.2. Metabolomics

Metabolomics aims to profile all the small molecule metabolites found within a cell, tissue (e.g. blood), organ and use this information to understand a dynamic response to a biological manipulation such as a pharmacological intervention. Metabolomics therefore offer the possibility of profiling large human populations or investigate a range of different tissues in animal studies both rapidly and cheaply or monitor metabolic responses to medicines. Thus, metabolomics has the potential to deliver a dynamic stratification of a patient population over time rather than a single snapshot measurement. This is important as we know from clinical validation studies that even if a patient population is selected, e.g. on the basis of a relevant genetic variant, often only a percentage of that patient population will respond to a particular treatment. For complex diseases the analysis of one (or a limited number of) genetic variant(s) can be beneficial in selecting patients for treatment, but is unusually not sufficient to achieve sensitivity and specificity values that approach 100% (i.e. all responders can be identified using the biomarker and non-responders are not selected for treatment). In this respect, techniques such as metabolomics and proteomics could potentially complement current methods of selecting patients for treatment (e.g. with targeted therapies).

However, there are a number of challenges related to these techniques which require an in depth understanding of the techniques and metabolic pathways. For example, some key pathways are better

represented by high concentration metabolites inside the cell, and thus, the coverage of the metabolome may become biased towards these pathways (e.g., the TCA cycle, amino acid metabolism). There is also the challenge of modelling datasets with large numbers of variables but relatively small sample sizes. The biases may also be different across different diseases. Similarly to proteomics, analysis of the metabolome can be either targeted or non-targeted: the non-targeted approach gathers data on all possible compounds in the sample, typically more than 4000 mass spectrometry peaks, while the targeted approach focuses on selected compounds, often to quantify changes thereof, for example in response to disease, or pharmacological treatment. Furthermore, imaging MS can provide information on the spatial distribution of metabolites from a thin tissue section, for example from an organ. A reconstructed two-dimensional image of the compounds provides the location of metabolites within the tissue.

NMR spectroscopy and MS are the main techniques used to probe the metabolome in biofluids and biopsies (Nicholson et al. 2012). The advantages of both techniques, NMR and MS, are that they can quantify small molecules with high accuracy and require only a small amount of sample.

MS is used to identify and to quantify metabolites and is more sensitive than many other methods (e.g. NMR) but has the disadvantage that it is destructive and requires significant sample preparation, using either chromatography or capillary electrophoresis, and as such is directly affected by the physical and chemical composition of the sample (Holmes et al. 2015). Thus, MS approaches tend to be less reproducible, more platform-dependent and susceptible to variability. Gas chromatography (GC) and liquid chromatography (LC) are the most commonly used techniques, particularly ultra-high-performance liquid chromatography (UPLC), is being used increasingly for metabolic profiling. Methods such as capillary electrophoresis (CE) allow the analysis of samples as small as a single cell. LC-MS is particularly suitable for large, thermo-unstable organic molecules including many secondary metabolites, larger carbohydrates, and lipids. However, surface-based mass analysis has emerged with new MS technologies focused on increasing sensitivity, minimising background, and reducing sample preparation.

However, the need to analyse metabolites directly from biofluids and tissues continues to be a challenge in current MS technology, largely because of the limits imposed by the complexity of the samples, which contain thousands to tens of thousands of metabolites of varying size, solubility and spatial distribution. Among the technologies being developed to address some of these challenges is Nanostructure-Initiator MS (NIMS), a desorption/ionization approach that does not require the application of matrix and thereby facilitates small-molecule (i.e., metabolite) identification. However, this does not provide spatial information. MALDI is an alternative technique but the application of a MALDI matrix can add significant background at <1000 Da that complicates analysis of the low-mass range (i.e., metabolites). In addition, the size of the resulting matrix crystals limits the spatial resolution that can be achieved in tissue imaging. Because of these limitations, several other matrix-free desorption/ionisation approaches have been applied to the analysis of biofluids and tissues.

SIMS was one of the first matrix-free desorption/ionisation approaches used to analyse metabolites from biological samples. SIMS uses a high-energy primary ion beam to desorb and generate secondary ions from a surface. The primary advantage of SIMS is its high spatial resolution (as small as 50 nm), a powerful characteristic for tissue imaging with MS. However, SIMS has yet to be readily applied to the analysis of biofluids and tissues because of its limited sensitivity at >500 Da and analyte fragmentation generated by the high-energy primary ion beam. DESI is a matrix-free technique for analysing biological samples that uses a charged solvent spray to desorb ions from a surface. Advantages of DESI are that no special surface is required, and the analysis is performed at ambient pressure with full access to the sample during acquisition. A limitation of DESI is spatial resolution because "focusing" the charged solvent spray is difficult. However, a recent development, LAESI is a promising approach to circumvent this limitation.

NMR-spectroscopy is the only detection technique which does not rely on separation of the analytes. It enables the measurement of all kinds of small molecule metabolites simultaneously. The main advantages of NMR are high analytical reproducibility with a detection limit in the sub-micromolar range, the simplicity of sample preparation and the non-destructive nature of the analysis. Moreover, new solid-state NMR methods allow analysis of samples of less than 1 mg, allowing to study tissue heterogeneity, e.g. between tumour tissue and tumour margins. Practically, however, it is relatively insensitive compared to mass spectrometry-based techniques. Other methods not as frequently applied, include ion-mobility-spectrometry, electrochemical detection (coupled to HPLC), Raman spectroscopy and radiolabelling (when combined with thin-layer chromatography).

Summary

The table below (Holmes et al. 2015) and Fig. 4 summarises the strengths and weaknesses of analytical platforms currently used in metabolic phenotyping

However, even when a single technique is considered there are multiple different commercial platforms available with variable measurement options which have different detection capabilities and hence differing bioanalytical performance (Holmes et al. 2015). In addition, there are likely to be differences depending on the type of sample/tissue chosen for analysis, e.g. a biomarker for a disease process in urine is unlikely to be the same as that in blood plasma. Any biomarker would then need to be validated against the current gold standard current diagnostic test (if available) to demonstrate e.g. specificity and sensitivity. Thus, there is a need for (i) standardisation of analytical techniques, and (ii) harmonisation of metabolic profiling and biomarker identification for clinical phenotyping in order to meet regulatory requirements. The Standard Metabolic Reporting Structures (SMRS) group, formed in 2003, recommends standards for conducting and reporting metabolomics studies (Lindon et al. 2005). Other groups have recognised the need for harmonizing procedures in the clinical and surgical setting (Nicholson et al. 2012).

Standardisation and harmonisation efforts are crucial in order to establish robust, reproducible and comparable diagnostic methods on different platforms. From the regulatory point of view the main question is, are there methods -running on different platforms- delivering comparable results with comparable clinical value?

Fig. 4: Strengths and weaknesses of analytical platforms used in metabolic phenotyping studies (Holmes et al. 2015)

Table 1 | Strengths and weaknesses of analytical platforms used in metabolic phenotyping studies

Platform	Method	Relative strengths	Relative weaknesses
Nuclear magnetic resonance (NMR) spectroscopy	Exploits the ability of spin active nuclei to absorb and re-emit pulsed electromagnetic radiation of a characteristic frequency pattern when placed in a magnetic field; interaction of nuclei with electromagnetic fields gives information about molecular structure, chemical environment and molecular motion	Highly reproducible Gives atom-centred connectivity information Low cost per sample (mainly reagent free) Exact quantification possible Detailed SOPs and experimental parameters available Minimal need for sample preparation, chemicals, reagents Relatively high throughput (10–15 min per sample) Good metabolite identification databases (HMDB,[152] BRMDB,[157] AMIX™ [Bruker BioSpin, USA]) 2D methods applied to multiple samples informs statistical spectroscopic analysis (to aid in metabolite identification) High linear dynamic range (~1 x 10^6) Nondestructive Analysis of wide range of chemical structures and molecular sizes Molecular compartment information via diffusional methods	Relatively insensitive (mitigated by use of cryoprobes and high magnetic fields) High capital cost of instrumentation Overlap of metabolites in 1D spectra (mitigated by increased magnetic field strengths and ≥2D methods)
Direct injection mass spectrometry (DI-MS)	Uses a nanospray source directly attached to a MS and does not require chromatographic separation; metabolites are identified and quantified using stable-isotope labelled standards	Rapid data acquisition (3 min for ionization runs) No crosscontamination or column carryover High sensitivity and stability Low cost analysis (no solvents or reagents) Ion fragmentation capabilities for metabolite identification Automated analysis Low sample volume requirement (~5–10 µL) Simultaneous metabolite quantitation and profiling Can use a variety of mass spectrometers of different configurations	Issues of specificity (no retention time information) Inability to separate isomers and isobaric species Spectral alignment can be challenging Differential ionisation with potential to cause errors in quantification (mitigated by stable-isotope labelled standards)
Ultra-performance liquid chromatography-mass spectrometry (UPLC-MS)	Uses chromatographic columns packed with small particles (1.7 µm) to allow the use of ultra high pressure elution with improved chromatographic separation and reproducibility	Profiling or targeted quantitative modes depending on the MS detector Sample handling simple (protein precipitation or dilution of biological sample) High throughput capability (typically 1–20 min per sample) UPLC can be coupled to any type of MS Any column chemistry possible, giving a wide range of detectable compounds	Retention times are highly specific to exact chromatographic conditions Databases only transferable when chromatographic conditions are identical Batch effects can be introduced by mass detector drift of chromatography Relatively young technology—metabolite databases are incomplete Variation between spectrometer systems
Gas chromatography-mass spectrometry (GC-MS)	Enables gas phase chromatographic separation of molecules followed by MS detection	High sensitivity High reproducibility Extensive public databases available for small molecule identification Mature technology	Only detects volatile compounds Run times relatively long (30–60 min) Samples require derivatisation, which can be time-consuming, and reagents, which can be costly Uses environmentally unfriendly solvents and reagents Multiple derivatisation of certain compound classes possible
Capillary electrophoresis-mass spectrometry (CE-MS)	Provides an electrokinetic separation method combined with a MS detector	Can be run in both polar profiling or targeted quantitative modes Excellent for polar analysis in aqueous samples Measures inorganic (Cl^-, SO_4^{2-}, NO_3^-) and organic anions ($C_2O_4^{2-}$) Low running costs Useful for the separation of proteins, peptides and metabolites	Not suitable for high throughput profiling due to system performance degradation Metabolite identification databases are not extensive
Rapid evaporative ionization-mass spectrometry (REI-MS)	Enables thermal disintegration of biological material using a modified surgical electrocautery device, followed by MS analysis of ionized metabolic constituents; used in surgical dissection for tissue identification	Real-time *in situ* analysis of intact tissues Rapid (0.1–3 sec analysis time) Chemical information can be mapped directly to histology Depth of information matches or exceeds *in vivo* labelling Can be used to analyse tissues, biofluids and other electroconducting materials	Destructive analyses Custom MS systems needed; still experimental Validation is difficult because of limited distribution of the technology (novelty of technique) Can produce noxious vapours
Desorption Electrospray Ionization MS	Separates molecules using DESI followed by MS analysis; rasters an analytical ion beam across tissue or other organic surfaces, desorbing ions that are measured by an atmospheric pressure inlet MS	Soft ionization technique, so facilitates analysis of fragile molecules Suitable for direct metabolite identification Untargeted multicomponent analysis Generates spatially resolved digital chemical information unlike histopathology Can work on untreated frozen sections DESI does not require background matrix for sample embedding Direct detection of topography, generation of multivariate images of tissue or organic surfaces	Slow scanning when used in high resolution mode (>1 h per sample) Limited availability of metabolic and/or lipidomic databases Detects only a selected range of metabolites (predominantly lipids)

Abbreviations: AMIX, analysis of mixtures; BMRDB, biological magnetic resonance databank; Cl^-, chloride; $C_2O_4^{2-}$, oxalate; DESI, desorption electrospray ionization; HMDB, human metabolome database; MS, mass spectrometry; NO_3^-, nitrate; SO_4^{2-}, sulphate; SOP, standard operating procedures.

5.1.3. Lipidomics

As for metabolomics, analysis of lipids faces multiple challenges due to the large structural complexity and heterogeneity in the compounds measured. Lipids can be divided into eight categories, i.e. fatty acyls, glycerolipids, sphingolipids, glycerophospholipids, saccharolipids, sterol lipids, prenol lipids and polyketides, which when combined with their complex structure creates significant analytical challenges (Sethi und Brietzke 2017). Lipidomics is broadly divided into two separate approaches: (1) a targeted, highly sensitive, and quantitative approach using liquid chromatography-tandem MS (LC-MS/MS), and (2) a high-throughput discovery (shotgun) lipidomics approach that may only detect the most abundant species (O'Donnell et al. 2014). Thus LC-MS/MS is the approach of choice when specific lipids are of interest and accurate quantitation is required; however, separation times can be long. Shotgun lipidomics enables the analysis of many samples in a short time but it is not quantitative and low abundance lipids are missed.

The development of sensitive bench top mass liquid chromatography-MS (LC-MS) has revolutionized the lipidomics field and enabled the study of small amounts of complex mixtures of diverse biological samples. O'Donnell et al (2014) highlights that LC-MS has solved many of the problems of the older approaches such as (1) low sensitivity and selectivity (thin layer chromatography and high-pressure liquid chromatography), (2) the need for time-consuming derivatization methods (eg, for gas chromatography/MS), and (3) the requirement for radioisotopes with their inherent health issues, notably in regard to ^{32}P-orthophosphate. When compared with this, LC-MS combines high sensitivity with the ability to detect, characterize, and quantify individual molecular species directly without derivatisation or purification (O'Donnell et al. 2014).

In addition, recent progress in mass spectrometric techniques such as ESI, DESI and MALDI, often in combination with separation techniques, have enabled an increase in the sensitivity, specificity and throughput of lipidomic screens from complex biological samples. However, none of the currently available MS methods are capable of outputting the complete lipidome and multiple MS platforms would be needed, significantly decreasing the analytical efficiency. NMR spectroscopy, fluorescence spectroscope and dual polarization interferometry techniques are also used; NMR has the advantage that it is quantitative and non-destructive, but it suffers from a lack of sensitivity and resolution of lipids in complex mixtures (Yang und Han 2016).

A summary of the various methodologies is provided in the table below (Fig. 5); none of the currently available MS are capable of outputting the complete lipidome; all approaches have inherent advantages and disadvantages which highlights the need to tailor the approach to the question asked. For example ESI-MS-based techniques offer the ability for rapid analysis potentially enabling the capture of metabolic changes driven by disease but may have a decreased sensitivity in the detection of low-abundant lipids and problems of quantitation in the absence of internal standards (Sethi und Brietzke 2017). More recently the combination of multiple approaches has enabled lipid imaging in tissue samples to determine the spatial localization of molecular species in an organ of interest, but the resolution of this technique is not yet at a cellular or subcellular level. As technological methods continue to evolve, a major challenge lies in the development of computational and bioinformatics tools for the analysis of the very large datasets produced.

Fig. 5: Advantages and Disadvantages of Analytic Approaches in Lipidomic Research (Sethi und Brietzke 2017)

Experimental Approach	Lipid Classes Covered	Advantages	Disadvantages
Chromatography			
Thin-layer Chromatography (TLC)	Solvent systems established for most lipid classes	Very established technique; technically relatively easy; does not require sophisticated instrumentation; spot chromatograms allow for rapid screening of mutant extract libraries.	Low resolution and sensitivity limits many lipidomic applications; detection of lipid by iodine vapour and (class-specific) dyes and radioactivity.
High-performance Liquid Chromatography (HPLC)	Many lipids, including sterols, GP, TG, DG, FA and lipid headgroup derivatives.	Well established with worked out reverse- and normal-phase conditions available; ease of automation; very quantitative.	Detection by refractive index or mass detector (lipids, as defined here, in general do not absorb visible and UV light effectively); medium sensitivities in general.
Gas Chromatography (GC)	Non-polar compounds such as TG; derivatized FA and sterols.	Very widely used for determination of fatty-acid composition, detection generally by MS.	Requires volatile compounds or derivatization of polar lipids.
Ultra-Performance Liquid Chromatography (UPLC)	Phospholipids, Triacylglycerides	Stable retention times and allows good separation even within the species of one lipid class; separation of isomeric structures and isobaric lipid species	Smaller particles generate greater back pressure, making it necessary to have a chromatography system
Mass spectrometry (MS)			
Atmospheric Pressure Chemical Ionization (APCI)	Non-polar lipids such as triacylglycerols, sterols, and fatty acid esters	High collision frequency; possible to use a nonpolar solvent as a mobile phase solution, instead of a polar solvent.	Low sensitivity for polar and ionic compounds; Limitations for the analysis of biopolymers, organometallics, ionic compounds and other labile analytes
Electrospray Ionization (ESI)	Polar compounds such as GP	Direct detection by m/z; high sensitivity and resolution; direct profiling of complex lipid mixtures; ease of automation; compatible with upfront LC separation, and requires minimal sample biomass.	Supression of ionization, in particular in the case of crude extracts and when low-abundance species are to analyzed; absolute quantification requires considerable efforts (for example, class and mass dependent internal standards).
Matrix Assisted Laser Desorption Ionization (MALDI)	Many lipids including complex glycolipids.	Direct detection by m/z; buffer and salt contaminants generally well tolerated; can be combined with prior TLC separation.	Supression of ionization, in particular in the case of crude extracts and when low-abundance species are to analyzed; matrix backgrounds.
Ion Mobility (IM)	Phospholipids, separation of complex lipid extracts from interfering isobaric species	Allows the determination of the collision cross section; provides a new set of hybrid fragmentation experiments; improves the peak capacity and signal-to-noise ratio of traditional analytical approaches	Limitation for precise determination of reduced mobilities (K_0) and collision cross sections (CCS)
Imaging	Glycolipids, Phospholipids, Neutral lipids	Determine the distribution of biological molecules by direct ionization and detection	Difficult to detect minor constituents or molecules that are not easy to ionize because numerous molecules exist in the crude mixture of tissue samples
Nuclear Magnetic Resonance (NMR)			
^{1}H	All lipids	Direct measurement; non-destructive; powerful technique for structural analysis of purified compounds.	Low sensitivity, spectra dominated by very abundant lipids (cholesterol, phosphatidylcholine).
^{31}P	Phospholipids	Direct measurement; non-destructive; quantitative	Line broadening of lipids in aqueous solutions; low sensitivity

5.2. Sources and Structure

Compared to genomics, public deposition and storage of MS-based proteomics data are still less developed due to the inherent complexity of the data and the variety of data types and experimental workflows (Perez-Riverol et al. 2015). This is even more true in the fields of metabolic and lipidomics, which are typically considered less matured than the field of proteomics.

Examples of relevant databases are provided in the following sections.

5.2.1. Proteomics data

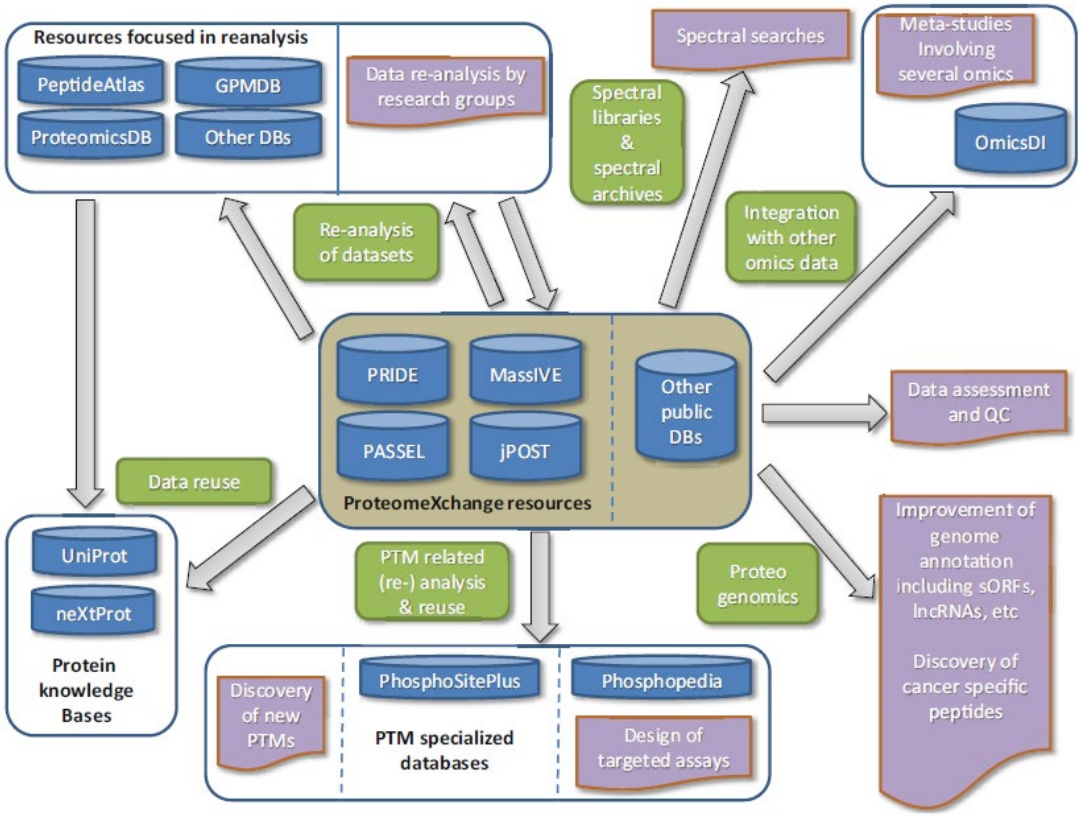
- Well established databases (see also Fig. 6) for proteomics data are the Global Proteome Machine Database (GPMDB), PeptideAtlas, and the PRIDE database. Additionally, other resources such as ProteomicsDB, MassIVE (Mass Spectrometry Interactive Virtual Environment), Chorus, MaxQB, PASSEL (PeptideAtlas SRM Experiment Library), MOPED (Model Organism Protein Expression Database), PaxDb, Human Proteinpedia, and the human proteome map (HPM). Furthermore, there are several more specialized resources.

- The SWISS-PROT is an annotated protein sequence database consists of sequence entries, which are updated monthly. Sequence entries are composed of different line types, each with their own format. For standardization purposes, the format of SWISS-PROT (see http://www.expasy.ch/txt/userman.txt) follows as closely as possible that of the EMBL Nucleotide Sequence Database. The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (i) annotations, (ii) minimal redundancy and (iii) integration with over 100 other databases. In 2004, all annotations of PTM features were standardized by creating a controlled vocabulary and updating the entries in Swiss-Prot (Farriol-Mathis et al. 2004). UNIPROT (Fig. 6) combines the reviewed and manually annotated SWISS-Prot database with the unreviewed and computationally analysed TrEMBL database.
- Recently, the ProteomeXchange (PX) consortium has been formed to enable a better integration of public repositories to implementation of standardized submission and dissemination pipelines for proteomics information. By August 2014, PRIDE, PeptideAtlas, PASSEL, and MassIVE are the active members of the consortium (Fig. 6).
- The SCALLOP consortium studies genetic associations with proteins as a collaborative framework on the Olink Proteomics platform.

Challenges related to data sources and structures:

- It is important to mention here that no single proteomics data resource will be ideally suited to all possible use cases and all potential users.
- The identification of false-positives and false-negatives in big resources or when different datasets are combined is still a challenge (Perez-Riverol et al. 2015). Before comparing resources, standard criteria should be set in order for the comparisons to be meaningful: (i) sample source; (ii) sample handling and LC-MS/MS technology; (iii) depth (comparable number of runs, technical sophistication; informatics must account for variations in depth; (iv) search libraries/databases; (v) search algorithms; (vi) modifications searched; (vii) mapping to proteins; (viii) protein interference (method for removing redundancy from the list or proteins with peptide evidence); (ix) error rate (result sets should have well-defined and low false discovery rates) (Farrah et al. 2014).
- Most of these databases accept data from heterogeneous sources, which presents a challenge in analysis. For instance, data acquired with different proteomics technologies, different computational pipelines and different quantification strategies may be combined in the database (Schaab et al. 2012), resulting in an uncontrolled false discovery rate.
- Several major standardisation initiatives worldwide are currently ongoing in order to provide information on assays that can be multiplexed, standardized, reproduced and shared across laboratories and instrument platforms. The ultimate aim is to drive a precise and highly specific quantification of proteins in a robust and reproducible manner (see section 5.4).
- Finally, another challenge is that at present, studies integrating different "omics" technologies are becoming more common. This type of studies poses a challenge for traditional repositories (which are usually field-specific) and researchers. However, it is far from straightforward to link data from different approaches, for instance MS-based proteomics and RNAseq data obtained in the same study. As discussed, validating, the associations between proteomics and other omics will be key to understanding their functional associations.

Fig. 6: Overview of the Main Uses and Applications of Public Proteomics Data Sets (Martens und Vizcaíno 2017).



5.2.2. Metabolomics data sources

In general, there is a limited availability and incompleteness of the databases to identify metabolites. Validation of findings is difficult because of the limited distribution and high costs of the technology and analytical variation between spectrometer systems. Examples of metabolomics data sources include:

- MetaboLights (http://www.ebi.ac.uk/metabolights). Database for Metabolomics experiments and derived information. The database is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.
- ChEBI (https://www.ebi.ac.uk/chebi). Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds. The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms. ChEBI incorporates an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children entities are specified.
- Metabolomics-Workbench: NIH funded; includes databases and reference standards; The Metabolomics Workbench Metabolite Database contains structures and annotations of biologically relevant metabolites. The database contains over 61,000 entries, collected from public repositories such as LIPID MAPS, ChEBI, HMDB, BMRB, PubChem, and KEGG. The Human Metabolome

Gene/Protein Database (MGP) of metabolome-related genes and proteins contains data for over 7300 genes and over 15,500 proteins. The main objective of RefMet is to provide a standardized reference nomenclature for both discrete metabolite structures and metabolite species identified by spectroscopic techniques in metabolomics experiments — an essential prerequisite for the ability to compare and contrast metabolite data across different experiments and studies. To this end, a list of 42,000 names from a set of over 200 MS and NMR studies on the Metabolomics Workbench has been used as a starting point to generate a highly curated analytical chemistry-centric list of common names for metabolite structures and isobaric (compounds with same nominal mass but different molecular formulae) species that present unique separation challenges.

5.2.3. Lipidomics data sources

Examples of lipidomics data sources include the following:

- LipidMaps (http://www.lipidmaps.org). The NIH funded multi-institutional effort was created in 2003 with the aim of identifying and quantitating, using a systems biology approach and sophisticated mass spectrometers, all of the major — and many minor — lipid species in mammalian cells, as well as to quantitate the changes in these species in response to perturbation, such as upon treatment. The goal is to better understand lipid metabolomics in diseases. They provide several resources, including MS standards and a new lipid classification system - the first internationally accepted lipid classification, nomenclature, and structural representation system suitable for the complex bioinformatics data basing required analysing the numerous molecular species of lipids.
- Metlin: https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage; Metlin is an MS/MS metabolite database including lipids.
- HMDB: http://www.hmdb.ca: a freely available electronic database containing detailed information about small molecule metabolites found in the human body. It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery and general education. The database is designed to contain or link three kinds of data: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. The database contains 114,110 metabolite entries including both water-soluble and lipid soluble metabolites as well as metabolites that would be regarded as either abundant (> 1 uM) or relatively rare (< 1 nM). Additionally, 5,702 protein sequences are linked to these metabolite entries.
- ELIFE: European Lipidomics Initiative: The primary aim of the 'European lipidomics initiative' (ELIFE) was to mobilise and organize key stakeholders, researchers and end-users in the area of metabolomics, especially lipidomics research, and to further define this field of research in terms of participants, scientific contents and strengths.
- LipidBlast: http://fiehnlab.ucdavis.edu/projects/lipidblast. An in-silico database that can be used to annotate and identify hundreds of lipids in plants, bacteria, algae, animals, humans and viruses.

In conclusion, as outlined in sections 5.2.1 - 5.2.3 and as shown in Fig. 6, the proteomics/metabolomics/lipidomics are structured, but inherently complex and heterogeneous due to the large variety of data types. For proteomics, there are different levels of processing, from raw data files to peptide identification, to protein knowledge-based information. Databases accept data from heterogeneous sources and represent repositories of diverse lists for protein identification. For instance, PeptideAtlas supports a non-redundant protein identification list of 1% FDR, whereas GPMDB computes a confidence value for each peptide and lists all peptide and protein identifications output (Farrah et al. 2014), It will also be necessary to standardise how data are formatted in order to link

the given analyte to defined metabolites or proteins and the quantity (Lindon et al. 2005). To ensure precision for metabolic analysis, it will be imperative to standards the terminology utilised to identify compounds and to adhere to existing naming conventions, such as IUPAC (http://www.sbcs.qmul.ac.uk/iupac/) or KEGG (http://www.genome.jp/kegg/). Also, the linkage to quantities will be variable, as e.g. absolute quantities require a specific unit, whereas relative quantities requires referring to the context in which the values were measured.

5.3. Veracity: Data quality and validation

The following sections will focus on proteomics rather than metabolomics or lipidomics, since in general the latter two fields are much less advanced compared with proteomics and thus application in regulatory processes is likely to be farther away for these fields. However, the majority of the issues raised are may be relevant for all across "omics" technologies.

Compared to genomics, proteomics is much more complicated and dynamic and thus, could face more complex challenges with regard to ensuring reliability of the generated data. Proteins occur in a wide range of concentrations in the body, making it extremely difficult to detect low abundance proteins in a complex biological matrix.

Membrane proteins constitute 30% of the typical proteome, which creates particular challenges due to their propensity to aggregate and precipitate in solution which confounds their analysis (as well as potentially the measurement of other compounds). The target residues for tryptic cleavage (i.e., lysine and arginine) are mainly absent in transmembrane helices and preferentially found in the hydrophilic part of these lipid bilayer-incorporate proteins (Chandramouli und Qian 2009). As such some techniques are not suitable for the analysis of membrane proteins necessitating the development of alternative techniques.

Serum Proteomics and Biomarker Discovery: more than 10,000 different proteins are present in the human serum and many of them are secreted or shed by cells during different physiological or pathological processes. Thus, serum offers opportunities for biomarker discovery but creates huge challenges, given the complexity of the proteins it contains and the considerable differences in their individual concentrations, ranging from several milligrams to less than one picogram per millilitre. The analytical challenge for biomarker discovery in, for example, selection of a target patient population, arises from the high variability in the concentration and state of modification of many human plasma proteins between different individuals (Chandramouli und Qian 2009) in addition to the dynamic nature of expression over both location of measurement and time. Thus, even if the correlation between serum levels and concentrations in relevant tissues is well established, there are concerns that variability in sampling could lead to unwanted variability (e.g. false positives or false negatives). These factors are likely to impact downstream challenges to identify biomarkers of consistent stability and reproducibility for regulatory decision making.

5.4. Reproducibility of raw data processing and concordance between processing methods (a specific challenge of bioanalytical omics approaches)

To accurately evaluate proteomics data, a question that needs to be asked is how efficient a protein extraction method is. It is important to know whether different methods for protein concentration measurements will generate different results due to different yield. In addition, since proteins have to be extracted from membrane, cytoplasm, nucleus, and other organelles for proteomic analyses, different methods are prone to have different outcomes. While in theory all proteins should be included in the analysis, this is often not the case in real experiments when diverse extraction methods with

different yield efficiencies are utilized even for the same type of tissue or cell samples. In addition, sample preparation will have significant effects on results, especially for unstable biomarkers.

Similar concerns exist and are maybe amplified for metabolomics and lipidomic analyses, which if anything is more variable and dynamic than proteomics analysis. The next section will describe the challenges with regards to variability and its influence on data quality and validation.

5.5. Variability: Data heterogeneity and standards

For a biomarker to be applicable in the clinical setting, measurements need to be reliable, and measurements need to be reproducible across methods/platforms and laboratories, and inadequate analytical validity ultimately affects the clinical utility of the biomarker test. One major factor affecting reproducibility of measurements is the simultaneous elution of many more peptides than can be measured by mass spectrometers. This causes stochastic differences between experiments due to data-dependent acquisition of tryptic peptides. Early large-scale shotgun proteomics analyses showed considerable variability between laboratories (Peng et al. 2003; Washburn et al. 2001), although reproducibility has been improved using Orbitrap mass spectrometers (Tabb et al. 2010). Targeted proteomics, in general, demonstrates increased reproducibility and repeatability compared with shotgun methods, although at the expense of data density and effectiveness (Domon und Aebersold 2010).

A report in 2009 by the HUPO Test Sample Working Group revealed that out of 27 laboratories which examined the same sample that consisted of 20 highly purified human proteins, only 7 laboratories reported all 20 proteins correctly (Bell et al. 2009). A subsequent centralized analysis of the raw data revealed that – nevertheless - all 20 proteins had in fact been detected by all 27 labs. The sources of problems encountered in the study included missed identifications (false negatives), environmental contamination, database matching, and curation of protein identifications. As in recent years search engines and databases improved, it is likely that the fidelity of MS-based proteomics has increased. However, in the context of regulatory decision making where such proteins may be used to stratify patients or monitor efficacy, such inconsistency is a major cause for concern.

This highlights the requirement for systematic benchmarking and validation of those analytical technologies, with particular focus on the reproducibility of results. Furthermore, there is a need to better understand variability and to agree on harmonised protocols.

To analyse the robustness, reproducibility and utility of a targeted proteomics approach, a specific strategy, the dimensionless retention time concept as an advance, iRTs, was examined in a multicentric setting, involving 28 laboratories (Vialas et al. 2017). This initiative revealed that transferring and sharing peptide retention times across different chromatographic set-ups both intra- and inter-laboratories is feasible. Parallel quantitative analyses showed a high reproducibility despite the variety of experimental strategies used and the diversity of analytical platforms employed. However the authors discuss that the highly reproducible results are based on the centralized preparation of the samples used in this study (Vialas et al. 2017). This suggests that harmonized protocols and analysis methods could help to increase reproducibility.

Several groups have evaluated the number of replicates necessary to observe a particular percentage of the proteins in a sample (Liu et al. 2004; Kislinger et al. 2005; Slebos et al. 2008). Others have examined how the numbers of spectra matched to proteins compared among analyses (Balgley et al. 2008; Washburn et al. 2003). The few comparisons across different laboratories for common samples have shown low reproducibility (Bell et al. 2009; Chamrad und Meyer 2005; Omenn et al. 2005) raising significant concerns with regard to reproducibility. Tabb et al., 2010, observe peptide lists from pairs of technical replicates overlapped by only up 60% at maximum (Tabb et al. 2010). Moreover, our

proteomics experts report when analysing a replicate from an identical sample, the approximately 3000 peptide sequences detected demonstrate a typical overlap of only approximately 60%. Several rounds of technical replicates could help to increase the statistical overlap and recovery rate.

The complexity of the analyses leads to variation in the peptides and proteins identified. Minor differences in liquid chromatography, for example, may change the elution order of peptides or alter which peptides are selected for MS/MS fragmentation. Small differences in fragmentation may cause some spectra to be misidentified by database search software. A major source of variability is the sampling of complex mixtures for MS/MS fragmentation. As a result, proteomic technologies may be inherently less amenable to standardisation than e.g. DNA sequencing. Existing literature has emphasized measurement variability in response to changes in analytical techniques: (i) fluctuations can occur in the autosampler, drawing peptides from sample vials (van Midwoud et al. 2007); (ii) the use of multidimensional separation can introduce more variation (Delmotte et al. 2009; Slebos et al. 2008); (iii) instrument platforms (Elias et al. 2005); identification algorithms ((Resing et al. 2004; Kapp et al. 2005), and many configuration choices during the bioinformatics steps of protein identification (Tabb 2013) produce variability (Tabb et al. 2010).

Several initiatives have created tools to minimize variability and increase standardisation. One example is bioinformatics tools such as Scaffold which increases the confidence in protein identification reports through the use of several statistical methods (Searle 2010) or several statistical assessments combined to examine the quality control (Wang 2017). Another example is a standardized AP-MS workflow that (i) helps to achieve interlaboratory reproducibility despite differences in mass spectrometry configurations and (ii) improves the sensitivity by combining independent data sets (Varjosalo et al. 2013). Besides standardisation efforts and novel proteomics technologies, comparative data analysis will improve data quality in proteomics (Mann 2009).

Analysis of variability in system performance entails two different measures that are often conflated. The first is repeatability, which represents variation in repeated measurements on the same sample and using the same system and operator. When analysing a particular sample, the same way on the same instrumentation, the variation in results from run-to-run can be used to estimate the repeatability of the analytical technique. The second is reproducibility, which is the variation observed for an analytical technique when operator, instrumentation, time, or location is changed. In proteomics, reproducibility could describe either the variations between two different instruments in the same laboratory or two instruments in completely different laboratories (Tabb et al. 2010).

A high quality proteomics dataset should fulfil at least two criteria with regards to variability: 1) repeatability of data generated within the same laboratory and 2) consistency of data generated from different laboratories (Gu und Yu 2014). Such variation has major implications for the use of proteomics to support regulatory decision-making. Therefore, standardisation efforts are essential.

Major standardisation efforts:

Proteomics standardization initiatives include the following:

- Proteomics Standards Initiative by the Human Proteome Organization (HUPO-PSI).

Recommendations for standardisation of collection, integration, storage, and dissemination of proteomics data were published. They aim at delivering a set of guidelines representing the minimal information required to report and sufficiently support assessment and interpretation of a proteomics experiment ("minimum information about a proteomics experiment;" the MIAPE guidelines). Additionally, the MIAPE – Mass Spectrometry Informatics module has been designed to specify a minimal set of information to document a mass spectrometry-based peptide and protein identification and characterisation experiment. The MIAPE: Mass Spectrometry Quantification module identifies the minimum information required to report the use of quantification techniques in a proteomics

experiment, sufficient to support both, the effective interpretation and assessment of the data and the potential recreation of the results of the data analysis. As for all MIAPE documents, these guidelines will evolve and will be made available on the PSI website at the url http://psidev.info (Taylor et al. 2007), The MIAPE initiative is embedded within the framework of The Minimum Information for Biological and Biomedical Investigations (MIBBI) project (Taylor et al. 2008).

- Clinical Proteomic Technology Assessment for the Cancer by National Cancer Institute.

The ProteomeXchange (www.proteomexchange.org) consortium has been set up to provide a single point of submission of MS proteomics data to the main existing proteomics repositories such as PRIDE and PeptideAtlas. Two new members have joined the consortium: MassIVE and jPOST. ProteomeCentral remains as the common data access portal, providing the ability to search for data sets in all participating PX resources, now with enhanced data visualization components. The primary goals of these initiatives were aimed at facilitating data comparisons from different laboratories and for overall data quality evaluation (Deutsch et al. 2017).

Harmonization efforts of the Clinical Proteomic Tumour Analysis Consortium (CPTAC) portal (http://assays.cancer.gov) serves as an open-source repository of well-characterized targeted proteomic assays. The goal is to enable robust quantification of all human proteins and to standardize the quantification of targeted MS-based assays to ultimately enable harmonization of results over time and across laboratories (Whiteaker et al. 2016).

Other current, global initiatives are aimed at the development of common resources of validated protein–protein interactions and protein-binding molecules for standardized characterisation of the human proteome, especially its potentially clinically relevant constituents. These programs include ProteomeBinders, AffinityProteome (both at www.proteomebinders.org), Affinomics (www.affinomics.org), Human Protein Atlas (www.proteinatlas.org), and Human Antibody Initiative (HUPO). A noteworthy aspect of these initiatives is their common focus on developing standard approaches and criteria for characterisation of affinity reagents for analysis of the human proteome (Stoevesandt und Taussig 2012; Gloriam et al. 2010; Taussig et al. 2007). Such criteria are virtually nonexistent despite the fact that antibody binding-based techniques are usually considered as the "gold standard" in specific qualitative detection of a protein (Ivanov et al. 2013).

Metabolomics Standardisation efforts include the following:

CIMR (http://msi-workgroups.sourceforge.net/): The Metabolomics Standards Initiative's Core Information for Metabolomics Reporting (CIMR) comprises modules for particular aspects of metabolomics workflows; various biological disciplines (for example, microbiology, mammalian biology, plant biology); analytical techniques such as chromatography and NMR; and the use of various statistical techniques.

Standardisation recommendation for linkage to clinical output

In order to link proteomics data to clinical outcome data and be able to combine such datasets, it is important not only to standardize proteomics datasets but also the clinical data. In this respect, clinical data standardisation efforts, such as CDICS (https://www.cdisc.org/), are also important. Standardisation efforts like CDISC would facilitate sharing of structured data across different systems. So far, two industry initiatives can be identified using CDISC standards: CSDR (https://www.clinicalstudydatarequest.com/Default.aspx) and the Yoda project (http://yoda.yale.edu/data-holders).

Standardisation of informatics technologies and data formats

One of the most important themes of current standardisation initiatives centers around the standardisation of applied informatics technologies. Several major directions in standardisation efforts can be currently identified, including bioinformatics frameworks for data sharing, integration, processing, and interpretation. However, a hurdle to these efforts are the proprietary data formats and data processing algorithms associated with commercial mass spectrometers, programming interfaces and software platforms for data acquisition and processing (Chambers et al. 2012). Availability of standardized, platform-independent computational pipelines for analysis and interpretation of proteomics data as well as establishment of common data output formats are essential for data sharing and unbiased benchmarking of technologies.

Considerable efforts have been made to generate common data formats (Deutsch 2008; Côté et al. 2010; Martens et al. 2011; Pedrioli et al. 2004; Schramm et al. 2012; Keller et al. 2005) and data format converters (Chambers et al. 2012; Keller et al. 2005; Kessner et al. 2008; Sturm et al. 2008). Development of standardized platforms will enable rigorous assessment of diverse computational approaches not only to accelerate proteomics research but also to harmonize data (Ivanov et al. 2013).

5.6. Velocity: Speed of change

Speed of change in proteomics techniques

A study in 2008 evaluated that until then around 109 unique protein serum markers were approved by the FDA including over 20 protein-based cancer biomarkers (Anderson 2010). Astonishingly, the majority of these (88 of 109) were developed before 1993 by immunoaffinity assays. After the advent of proteomics (around 1995; (Wilkins 2009)) only 22 additional diagnostic protein markers were approved (Steffen et al. 2016). Thus, on average the speed from discovery to approval has been extremely slow with 1.5 new approved protein assays by the FDA per year. This suggests that the current discovery pipeline has been inefficient and potentially suffering from high false-positive rates, which hampers the validation of true biomarkers (Vidova und Spacil 2017). Robust assays that adhere to precision, repeatability, reproducibility, linearity, limits of quantification, matrix effects, selectivity, and analyte stability are needed to accelerate the biomarker discovery pipeline.

The genome is relatively static, whereas the proteome, metabolome and lipidome are dynamic and spatial. Proteins are continually undergoing changes, e.g., binding to the cell membrane, interacting with partner proteins to form protein complexes, or undergoing synthesis and degradation. Similarly, metabolites and lipids by definition have a signalling role and their formation and destruction, and modification may be highly dynamic.

The proteome is highly variable. The measured variation is considerable from person to person under different environmental conditions, across different tissues or even within the same person at different ages or health status due to e.g. post-translational changes or changes in protein-protein-RNA-complexes that may vary and reflect functional changes. These challenges equally apply to the metabolome and lipidome. Thus, developing biomarkers which meet the criteria set out in section 5.1 is highly challenging. Proteomics data change from condition to condition, and over time as a result of the analysis or the disease process. Moreover, the comparability to genomics is not always a given. For example, the level of transcription of a gene gives only a rough estimate of its level of translation into a protein. An mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in only a small amount of protein.

5.7. Accessibility of data

Most databases described above include data from publicly research initiatives. The public availability of relevant 'omics' data via appropriately accessible registries and databases is a key requirement for developing Big Data analysis approaches.

An important aspect with significant impact on the accessibility and usability of data are related to privacy law, i.e. in case of analysing data derived and linked to participants in clinical studies or individual patients, an appropriate protection of personal data has to be ensured. There are differences in privacy law between different countries and regions that have to be taken into account in time (as early as possible) and addressed within specific projects/studies or efforts to make these data available for secondary analysis/data sharing.

Approaches for the management of patient level data which do not require the physical transfer from data like DataSHIELD might be helpful in this context (Budin-Ljøsne et al. 2015).

5.8. European guidelines on regulatory use of 'bioanalytical omics' data

No specific EMA guidelines are available for proteomics, metabolomics or lipidomics technologies. The EMA guideline on bioanalytical method validation EMEA/CHMP/EWP/192217/2009[1] entails recommendations for the validation of bioanalytical methods which can be applied to proteomics approaches. Refer to the part of this report on 'Genomics' for a more detailed summary of available guidelines in relation to omics and biomarkers.

5.9. General overview - method qualification

General methodical approaches and requirements for regulatory decision-making regarding biomarkers are specified in the guideline *Qualification of novel methodologies for drug development* (EMA/CHMP/SAWP/72894/2008)[2].

An overview of common challenges is provided in the guidance document 'Essential considerations for successful qualification of novel methodologies'[3]. The most relevant points to consider from this document are summarized below.

1. Definition of the Context(s) of Use (CoU)

- The Context of Use definition should take into account the (vi) Dependence to environmental conditions; the conditions will have a strong impact specifically when utilizing metabolomics approaches for biomarker discovery; and the (viii) Clinically feasible, e.g. biomarker status should be possible to determine with minimum discomfort for the patient
- The Context of Use is the critical reference point for the regulatory assessment of any qualification application.

2. Selection of Endpoint(s)

- (ii) Mechanistic understanding of the relationship between the biomarker, the disease and/or response to treatment; diagnostic and prognostic performance (sensitivity and specificity) should be demonstrated.
- (iii) Diagnostic accuracy: specific, sensitive and predictive (correlated with the disease state or

[1] http://www.ema.europa.eu/ema/index.jsp?curl=pages/includes/document/document_detail.jsp?webContentId=WC500109686%26mid=WC0b01ac058009a3dc
[2] http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004201.pdf
[3] http://www.ema.europa.eu/docs/en_GB/document_library/Other/2017/12/WC500239928.pdf

endpoint they wish to capture); predictive values for medicine response and likelihood ratio (LR) are to be addressed. Levels of positive or negative predictive value are to be characterised.

3. Statistical Analysis Plan (SAP)

- The study design and data analysis must support the intended context of use (CoU).
- The statistical planning of the qualification approach must be appropriate and follow a pre-specified statistical analysis plan (SAP).
- Exploratory studies and approaches can be included as appropriate.
- Generally, confirmatory studies and data sets are required to achieve the qualification of a method. It will have to be justified whether the qualification objective can be supported appropriately by retrospective studies, or whether prospective studies are required.
- Cross-validation approaches (e.g. due to limitations to obtain appropriately sized exploratory and confirmatory data sets in rare disease scenarios) should apply appropriate methodology and should be pre-specified, not envisaged *post hoc*.

4. Demonstration of clinical utility

- The impact on diagnostic thinking, patient management and clinical outcome must be specified and justified.

5. Standard of truth / surrogate standard of truth

- In case an assessment of the standard of truth (true status of the patient or the value of the measurement) is not possible or unethical (e.g. a required measure is too invasive), a surrogate standard of truth must be established and justified.

6. Appropriateness of the analytical platform

- The analytical platform, as intended for use, must be validated, (vii) clearly define thresholds for clinical decision points.
- Technical and performance characteristics must be specified and justified with respect to the Context of Use (CoU), including (iv) limits of detection/quantitation.
- The diagnostic performance must be (v) stable and reproducible across patients and across time.

Also, epidemiologic aspects may become important, for instance (i) Knowledge of the normal distribution of a biomarker across different demographics.

At the current time, few of these criteria are fulfilled by current proteomic, metabolomic or lipidomic biomarkers. Increasingly challenging from a regulatory perspective is the trend towards biomarker panels for identification of relevant patient populations. This includes determining the clinical value of the measured biomarker. This is partly driven by the difficulty of identifying a single biomarker of sufficient specificity and sensitivity for clinical utility. How will the relative components of each of the biomarkers be assessed? How will the relative changes over time of the individual components be understood? Maybe less challenging, but important is, to ensure robust and reliable diagnostic performance to guaranty the consistency of analytical results across laboratories.

EMA qualification advice[4] for novel technologies is a process by which applicants can obtain a CHMP opinion on the acceptability of a specific use of a method, such as the use of a novel biomarker. The link below provides details of some recent advice procedures, which include examples like plasma

[4] http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000319.jsp

fibrinogen as a medicine's development tool (prognostic biomarker), to identify COPD subjects at high risk for all-cause mortality or COPD exacerbations for inclusion in interventional clinical trials. Many of the issues raised during this advice procedure illustrate the complexities and challenges for the identification of truly prognostic biomarkers.

http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000319.jsp

5.10. Bioinformatics, algorithms, modelling and statistics

Bioanalytical Big Data analysis (i.e. models and algorithms) are subject to the growing field of data science which combines methods from various disciplines such as biostatistics, mathematical modelling and simulation, bioinformatics and computer science including data-integration, machine learning and high-performance computing.

The workflow of Big Data analysis approaches can be generally divided in two steps:

Big Data analyses are typically preceded ***Data Processing*** to integrate data from different sources followed by ***Data Analysis & Interpretation***: Based on these pre-processed (and curated) data inferences can be drawn using methods from biostatistics, mathematical modelling and simulation, bioinformatics and machine learning. While data processing can largely be automated, model and method development, validation and interpretation strongly rely on expert knowledge of data scientists.

Initiatives, such as Elixir, provide open label bioinformatics applications addressing numerous scientific problems and tasks. Thus, for instance, for proteomics, Elixir offers within its data platform bioinformatics applications addressing metadata documentation, standardisation, annotation and data management. The "tools platform" offers open user-friendly quantification algorithms and tools, along with direct coupling to dedicated and performant statistical analysis. The "interoperability platform" supports "multi-omics" approaches. The "compute platform" includes workflow analysis pipelines and activities related to the development of cloud infrastructures. Further important integrative activities are training and the quality control (QC) activities which are essential for both, data processing as well as for data analysis & interpretation. QC has historically not been as well developed in proteomics as in, for instance, the more established small molecule mass spectrometry. Especially for regulatory use within medicines development, QC activities are substantial to ensure safety and efficacy of the particular medicinal product or treatment strategies. Elixir is focused on activities to develop automatic and reliable pipelines for QC of proteomics data at different levels.

The analysis and biological interpretation of proteomics and other omics data can be challenging. Systems biology develops innovative network-based methods, which allow an integrated analysis e.g. of protein-protein interactions using e.g. functional similarity measurements of interacting proteins reflecting biological processes of investigated cells (Boyanova et al. 2014). Other approaches apply functional enrichment analyses to provide biological context within a protein network or densely connected subnetworks (Molecular Complex Detection (MCODE) analysis) and use the accumulative hypergeometric probability function to assign statistical significance to functional categories (Tripathi et al. 2015). An integrated network approach not only allows a detailed analysis of proteome networks but also yields a functional decomposition of complex proteomic data sets and thereby provides deeper insights into the underlying cellular processes of the investigated system.

By applying Big Data strategies, initiatives like Virtual Liver Network (https://fair-dom.org/partners/virtual-liver-network-vln/) or the PlateletWeb

(http://plateletweb.bioapps.biozentrum.uni-wuerzburg.de/plateletweb.php) aim to achieve insights in biological pathways and (patho-) physiological processes on the level of cells, organs or physiological pathways and functions, e.g. the immune system.

For the development of pharmacological models, the Drug Disease Model Resources (DDMoRe) consortium has been created to improve the quality, efficiency and cost effectiveness of Model-Informed Drug Discovery & Development (http://www.ddmore.eu). The DDMoRe Model Repository is built using the Pharmacometrics Markup Language (PharmML) and supports the Model Description Language (MDL) in order to facilitate the collaborative development of computational models. This is in line with efforts to develop standardised model descriptions such as, for example, through the systems biology markup language (SMBL) to allow for a databases of curated biological models (http://sbml.org/Main_Page; http://www.ebi.ac.uk/biomodels-main/).

Statistical and bioinformatics methods complement each other in data sciences. Thus, efficient linkage between bioinformatics and biostatistics is needed for an appropriate assessment of methods and approaches if applied in medicines development.

In case, "Big Data" analysis is used for regulatory purpose, the same requirements as for other (biostatistical) analytical methods and approaches are applicable, depending on the intended use.

Adaptations may be necessary to fully integrate these new technologies in regulatory guidance and processes. However, general requirements for the qualification and validation of novel methodologies for medicine development are defined as described in section 5.9.

6. Key case study illustrating regulatory challenges

The following key case studies were chosen as they represent relevant innovative examples of proteomics big data approaches using algorithm-driven big data technologies. These approaches are far advanced and have impact directly on the quality, safety and efficacy of the final medicinal product. Therefore, they are of importance in regulatory assessment.

6.1. Key case study: Active personalisation in therapeutic cancer vaccination utilizing proteomics approaches

Personalisation in the area of cancer immunotherapy includes stratification approaches based on the presence of specific genomic markers that predict the clinical response of patients toward certain therapies such as tyrosine kinase inhibitors or monoclonal antibodies. Another strategy for personalised cancer immunotherapy uses therapeutic vaccination approaches such as autologous tumour lysates or tumour lysate-loaded dendritic cells, where the actual antigens are not known. Though, in fact each patient is treated with an individual product this approach has been defined as passive personalisation since the actual antigens remained unknown. In contrast, **active personalisation** in therapeutic cancer vaccination relies on the manufacture of an individual vaccine for each patient with clearly defined antigens (Britten et al. 2013). Such cancer vaccines might consist of a cocktail of cancer-associated antigens, which are non-mutated proteins that are either over-expressed in tumour tissues or that are specifically expressed in tumours, while not expressed in most normal tissues. One approach to establish such kind of cancer vaccinations is based on synthetic peptides from such well-known tumour-associated antigens that can be manufactured in advance and can be stored as a so-called warehouse. Depending on the tumour-associated antigen expression profile (proteomics; measured by MS) an individual vaccine cocktail derived from the warehouse can be assembled for each patient. Translation of the APVAC concept into the clinical setting has now commenced in several consortia-driven projects recently announced: GAPVAC (http://www.gapvac.eu)

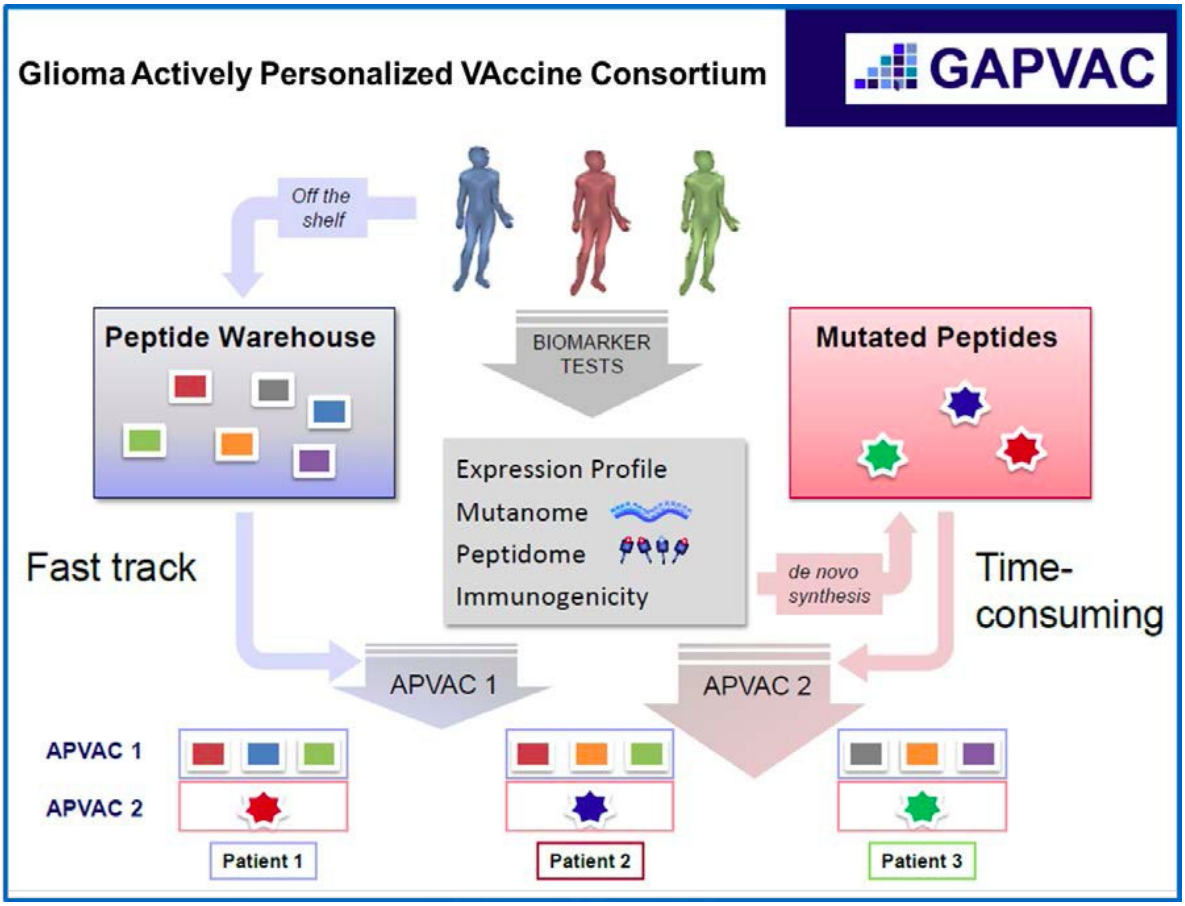
(see Fig. 7), MERIT (http://www.merit-consortium.eu) and HEPAVAC (http://www.hepavac.eu) as outlined the key case studies selected below.

6.2. Key case study: HLA ligandome analysis of tumours utilising proteomics (ligandome)

During recent years, mass spectrometry proteomics approaches have been developed to allow the isolation and identification of the HLA "ligandome" of tumour tissues (Rammensee und Singh-Jasuja 2013). Thus, isolation and characterization of all Class I and Class II binding peptides in a given cancer tissue sample has become possible. Upon comparison of cancer tissue ligandomes with healthy tissue ligandomes, it became evident that certain peptides are overexpressed or exclusively presented in tumour tissue, and not on normal tissue (Kowalewski et al. 2015). Due to this cancer-specific presentation on HLA molecules such peptides are an ideal basis for the manufacture of personalised anti-cancer immunotherapies. A warehouse of peptides identified in this way can be established and patients be treated with personalised peptide cocktails derived from the warehouse. Depending on each patient's specific HLA haplotype an individual peptide vaccination cocktail can be assembled.

Mass spectrometry also allows the direct identification of HLA-presented neo-epitopes in individual cancer patients. Such neo-epitopes arise due to tumour-specific mutations in the coding regions of genes. In contrast to the identification of mutations by sequencing approaches (see below) the selection of peptides by MS relies on the actual presentation of epitopes on HLA molecules. Neo-epitopes identified by MS can then be chemically synthesized and used as vaccines.

Fig. 7: Workflow of the GAPVAC study; credit: Prof. Dr. Stefan Stevanović, Tübingen, Germany



6.3. Application of bioinformatics tools in key case study: Active personalisation in therapeutic cancer vaccination combining genomics and proteomics or "pure" epitope prediction by bioinformatics algorithm prioritizing the epitopes

Another active personalisation approach relies on the identification of tumour-specific mutations (genomics, so-called "mutanome") that are used to manufacture peptide or mRNA cocktails encoding epitopes harboring the mutations. Epitopes derived from mutations are regarded as potential targets for cancer immunotherapy. They are called neo-epitopes as they lack expression in any healthy tissue. The systematic use of mutations for vaccine approaches, however, is hampered by the uniqueness of the repertoire of mutations (the mutanome) in every patient's tumour. Recently, a personalised immunotherapy approach was introduced targeting the spectrum of individual mutations.

'Bioinformatics tools are used for: mutation detection by exome sequencing, selection of vaccine targets by solely bioinformatical prioritization of mutated epitopes predicted to be abundantly expressed and good MHC class I or class II binders; followed by rapid production of synthetic mRNA vaccines encoding multiple of these mutated epitopes (NCT02035956).

The selection relies here based on bioinformatics algorithms as HLA ligands encompassing mutations that are not present in autologous benign tissue (neoepitopes) can be defined either by direct identification through LC-MS/MS or by epitope prediction.

6.4. Challenges in the use of proteomics approaches for personalised medicine

The key case studies described above illustrate the challenges that need to be overcome when using proteomics approaches for personalised medicine.

1) Bioanalytical method validation

A major regulatory question when using MS-based diagnostics to identify e.g. the ligandome of a cancer tissue is to what extent MS as a method can sufficiently be validated for bioanalytical method performance metrics such as accuracy, precision, specificity, sensitivity etc. It appears that currently it is not possible to sufficiently validate MS-based proteomics bioanalytical methods, according to information available. This is considered a major hurdle in regard to the potential for MS-based proteomics approaches to be usable in regulatory processes especially when there is a relevant impact on regulatory decision-making. It appears therefore, that for the time being alternative (non-proteomics) approaches to measure proteins are preferred.

2) Definition of standards for databases

For a broad application of personalised approaches, public databases are important, for example including information related to the ligandome profile of healthy tissues that can be compared to a patient's tumour ligandome (or any other profile comparing tumour to healthy tissue). From a regulatory point of view, it is important to define the requirements or standards to which such databases should adhere, such as common data formats (e.g. CDISC) or the qualification procedures standards for biomarkers (CPI, see section 8.5).

3) Regulatory challenges specific to bioanalytical Omics-based personalised approaches

Other regulatory challenges emanate from omics-based personalised approaches. The current paradigm of medicinal product development starts establishing a robust manufacturing process for a

single, well-defined product, followed by non-clinical studies and testing of the product in first clinical studies. Manufacture of individual peptide cocktails requires the establishment of a platform allowing the robust and consistent production of each drug product within the shortest period of time possible. The feasibility of such high-throughput manufacture will rely on flexible regulatory approaches such as adapted quality testing. Examples are stability testing and microbial testing which are time-consuming and need to be shortened. Preclinical pharmacological and safety testing is similarly affected. Active personalisation approaches and the resulting regulatory challenges have been discussed at EMA and have generally been viewed by EMA experts as a promising approach (Britten et al. 2013), Briefing Meeting Report Actively Personalised Vaccines (APVACs), EMA/88358/2012.

7. Applicability and Usability

Exploratory research

As described in previous sections, the usability of proteomics/metabolomics/lipidomics technologies in regulatory processes and in particular for regulatory decision-making is currently low as a result of important challenges that have to be overcome e.g. in relation to bioanalytical method validation and issues with variability in the acquired data and reproducibility of measurements. On the other hand, these technologies have a unique potential to establish a linkage between 'biological' physiological measures and clinical outcome. Many of the current and potential applications may lie in exploratory research, e.g. Biological pathway and target discovery or hypothesis generation based on causal associations between 'omics' profiles and signatures and consequential phenotypic outcomes in both health and disease and eventually link to drug responsiveness.

Thus, currently the profiling investigations are focused on whole exome sequencing and/or messenger RNA sequencing but ultimately down the line may progress to examining protein, metabolomic or lipidomic signatures. However, as discussed above, methodological challenges like variability and complexity of relevant biological pathways remains a cause of concern, and consequently there is increasing interest in e.g. refining accuracy by using panels of biomarkers as opposed to a single biomarker; for example a panel of proteomic biomarkers which includes markers of inflammation (IL-6, TNF-α) and mineral and bone disorder biomarkers (OPG, OPN, OCN, FGF-23, and Fetuin-A), was found to be potentially more relevant than a single biomarker to detect early stages of chronic kidney disease (Mihai et al. 2016). This exemplifies that panels of biomarkers, in future, very likely identified by Bioanalytical Omics technologies selected by specific algorithm and modelling might be of highly relevant use in clinical settings.

For further considerations addressing the underlying biological complexity for better understanding of (patho-)physiological pathways and the result of protein biomarker results see below under section *Patient stratification*.

Digital Repositories (e.g expression Atlases)

Established databases for proteomics data (see section 5.2.1. Proteomics data) are, for instance, the Global Proteome Machine Database (GPMDB), PeptideAtlas or the PRIDE database. These databases (expression atlases) of expression profiles of functional proteins like receptors or surface proteins in specific organs and tissues of humans and animals used in in-vivo models allow predictions of pharmacological, toxicological/adverse and immunologic effects in humans and predictions of interspecies comparability. As an example, a relevant initiative, again in the field of oncology, is the **Human Pathology Atlas**, a part of the Human Protein Atlas program (Uhlen et al. 2017) in which the global expression patterns of all protein coding genes in 33 cancer types for 9666 patients for whom clinical metadata existed, were compared with gene expression patterns in normal healthy tissues obtained from 162 individuals. What is perhaps more noteworthy is that when comparing across

different cancer types and location, 2375 of the prognostic genes showed opposite effects on prognosis depending on the cancer type and location **highlighting the need to perform functional studies of prognostic genes**. Moreover, as increasingly becoming recognised, there is an overlap of prognostic genes, both favourable and unfavourable, across some cancer types, for example renal, breast, lung and some pancreatic cancers. From a regulatory perspective, such studies may mean that in the future drugs will no longer be indicated for a particular cancer based on its location but for a particular oncogenic marker irrespective of the origin of the tumour.

Patient stratification

'Omics' technologies also contribute to improvements in the (early) stages of medicines development and are in use in clinical practice.

There are two main objectives for the application of patient stratification approaches.

1) Improving and accelerating clinical medicines development e.g. i) for potential prediction of disease progression; ii) for risk stratification or as part of an enrichment strategy in a randomised trial; iii) as a surrogate marker for a "hard" clinical endpoint.

2) In clinical practice, diagnostic stratification enables more specific and efficacious treatment for individual patients and clinical decision-making.

In particular, the potential to accurately stratify a patient population allowing to predict the downstream phenotypic is a prerequisite to establish individualised / precision medicines concepts.

For example, a number of PD-1/PD-L1 inhibitors are approved including companion diagnostics using PD-L1 immunohistochemistry protein biomarker assays. Several studies examining the usefulness of PD-L1 IHC assays have demonstrated a direct correlation of response rate to tumour PD-L1 expression level. However, the establishment of cutoff values is essential for appropriate clinical decision-making. It is required to specifying the PD-L1 expression (measured with a specific assay) defining the expression level the clinical benefit/risk balance is positive for the administration of a particular PD-1/PD-L1 inhibitor in a specific indication. The cutoff values for these assays vary from as low as 1% to as high as 50%. To allow for comparison, sensitivity and specificity of the bioassay for a given malignancy were calculated based on the reported objective response rate in individuals who were considered to have PD-L1 positive tumours (ORR+) and that of the individuals who were considered to have PD-L1 negative tumours (ORR-). The sensitivity and specificity of the IHC assays are generally poor. The stringent application of the results of these assays may potentially exclude individuals who may benefit from treatment or include patients who may not benefit. Also the interchangeability of the current assays is likely to be a challenge, which is a regulatory challenge (Diggs und Hsueh 2017).

The reasons for this observed low sensitivity and specificity example companion diagnostics approach are multifarious. Bio-analytical challenges, as discussed above, are contributing but have likely a generally minor impact. The mechanism of PD-L1 expression is complex. Multiple factors, intrinsic and extrinsic, appear to influence both, PD-L1 expression and response to treatment (Diggs und Hsueh 2017). For instance, specific cancer mutations (e.g. BRAF mutations in melanoma) contribute to altered response to treatment or even altered expression of PD-L1 (e.g. mutation in KRAS in the tumour). Moreover, extrinsic factors like cigarette smoking or platinum-based chemotherapy can increase PD-L1 expression (Diggs und Hsueh 2017).

This example underlines the limitation of current approaches driven on specific (single) biomarkers. New concepts, like Big Data driven analysis approaches, are required to address the underlying biological complexity to improve patient stratification, either for research and development of medicinal products or for an improved use of medicines in the clinical practice.

Design and manufacturing of personalised therapies

As outlined in the key case study (section 6) *Active personalisation in therapeutic cancer vaccination utilizing proteomics approaches*, 'bioanalytical omics ' driven Big Data analysis approaches can be an essential tool for the selection, optimisation or design of specific active agents used and applied for medicinal products developed as personalised therapies. The quality as well as the efficacy and safety of those products is directly dependent on the accurate and reliable function of the underlying 'bioanalytical omics' driven Big Data analysis approaches, therefore, the regulatory system has to follow the development of these new concepts and to prepare for future applications for marketing approvals for those kind of innovative treatments, see recommendations below.

Areas of current/potential applications	Benefits
Exploratory research	'Omic technologies' are an indispensable tool in basic biomedicinal research and are used in a broad range of research approaches like biological pathway and target discovery or hypothesis generation based on causal associations with phenotypic outcomes
Digital Repositories (e.g expression Atlases)	Allows 'in silico' predictions of pharmalogical, toxicological/adverse and immunologic effects in humans and prediction of interspecies comparability.
Patient stratification	Two main objectives: Improving and accelerating clinical medicines development e.g. i) for potential prediction of disease progression; ii) for risk stratification or as part of an enrichment strategy in a randomized trial; iii) as a surrogate marker for a "true" clinical endpoint. Diagnostic stratification in order to enable more specific and efficacious treatment for individual patients and clinical decision-making
Design and manufacturing of personalised therapies	Key case study (section 6): Active personalisation in therapeutic cancer vaccination utilizing proteomics approaches

8. Conclusions

The inherent complexity of proteomics and bioanalytical 'omics technologies' is a general challenge, which has also impact on the use of proteomics in the regulatory context. Comprehensively

characterising an entire proteome, for example, still poses a considerable challenge for the field. While the genome is relatively static, the proteome, as well as the metabolome and lipidome, are highly dynamic and spatial. Proteins are continually undergoing changes, e.g., binding to the cell membrane, interacting with partner proteins to form protein complexes, or undergoing synthesis and degradation. As a result, variability in proteomics sample measurements is considerable from person to person, across different tissues, or even within the same person.

Bioanalytical omics technologies in principle hold great promise. Meaningful changes at the proteomic level are not always present at the genomic level and therefore the proteome may be a better predictor of physiology, especially in disease. In combination with other approaches 'omics" technologies, namely metabolomics and lipidomics, can deliver a dynamic stratification of patient populations over time rather than a single snapshot measurement.

There is a unique potential that 'omics technologies' foster and accelerate the development of new treatments. Thus, the development of these technologies should be supported also by the regulatory system by addressing of general and specific challenges listed below in order to establish a robust framework for innovative therapy development enabling the full benefit for patients.

8.1. Bioanalytical method validation

The establishment of appropriate method validation of bioanalytical methods used in 'bioanalytical omics' approaches are required, in order to ensure reproducibility, sensitivity and specificity.

However, approaches sufficiently tailored to the context of 'bioanalytical omics' techniques may be required, due to the complexity of the techniques used to measure.

The current framework for bioanalytical (technical) method validation (laid out in the guidelines) provides the basic principles. Appropriate method validation approaches sufficiently tailored to the context of 'bioanalytical omics' techniques taking into account the complexity of the techniques used have to be established in order to ensure quality of results when used in regulatory processes.

Clear guidance should be provided for the bioanalytical method validation in the context of Big Data `bioanalytical omics` applications, depending on the impact on diagnostic, medical or regulatory decisions.

8.2. Comprehensiveness of available data sets

According to the 2nd recommendation of the FAIR Data Action Plan, The Open Data Mandate for (publicly funded) research should be made explicit in all policy. It is important that the maxim 'as Open as possible, as closed as necessary' be applied proportionately to use the genuine best efforts to share.

A major challenge is the availability of relevant data sets to carry out comprehensive analyses of (Big) 'omics' data. Commonly, starting with variability in sampling up to limitations of analytical methods, e.g. the stochastic differences between experiments inherent to proteomics, all these factors are affecting reproducibility and variability and as consequence, the comprehensiveness of data sets (see sections 5.3 – 5.5.). In particular in proteomics, there are significant challenges to obtain comprehensive data sets as the complexity of proteomics analyses leads to variation in the peptides and proteins identified (5.1.1 and 5.5). On top of the limited availability of data due to methodical analytical reasons, legal, ethical and commercial limitations exacerbate the problem.

A major challenge is the availability of relevant data sets to carry out comprehensive analyses of (Big) 'omics' data. Missing data / information may either limit capabilities to identify existing associations and pathways or may introduce bias leading to erroneous results and conclusions.

Depending on the specific development programme or the scientific questions addressed, the impact of these methodical limitations is variable. An adequate analysis of the completeness of available date and the potential impact of missing data / information has to be implemented timely during the development of 'omics technologies' projects.

Specific guidance should be provided how the completeness of analysis, in particular the missing data / information has to be addressed in Big Data 'omics' approaches, depending on regulatory purpose the results are intended to be used.

8.3. Data Quality

A crucial challenge associated with the use of proteomics and associated techniques in the regulatory and specifically the big data setting is the issue of controlling data quality. Control over data quality is critical on many levels in the case of proteomics data and will likely be more difficult than with other techniques (e.g. genomics). In addition to the issue of controlling the quality of the bioanalytical method itself (5.3.-5.4), the sample quality is also of crucial importance e.g. in proteomics analyses. Specifically, in big data analyses, samples will be derived from a variety of sources (e.g. because it concerns secondary use of existing samples). If low quality samples are used, the reliability of the data can become jeopardised. Also, prior handling of data (data analysis) may potentially be difficult to control from quality perspective, for example when secondary databases are used which comprise data derived from the results of analysing primary data.

The usability of 'omics technologies' in the regulatory context is dependent on standardisation of quality control methodologies on all levels of the analyses, from sample quality to data analysis methodologies.

The quality of the samples and the results of analysis have to be documented by providing relevant quality attributes via appropriate *data (file) formats* and *data standards*. Aspects relevant to quality of bioanalytical method validation as well as *Data Processing* and *Data Analysis & Interpretation* are addressed in the specific recommendations.

It will be in the remit of regulatory networks, in Europe and globally, to define which quality levels and attributes will be acceptable for which regulatory purpose, depending on the impact on diagnostic, medical or regulatory decisions.

8.4. Supporting the harmonisation and sharing of data (file) formats (standard open file formats)

A challenge specifically predominant in the field of "bioanalytical omics" is the management of heterogeneous data types and (file) formats. Defining harmonised data formats will be an important objective in relation to use of 'omics' data in the regulatory context as interoperability and availability of comprehensive data sets (see above) are crucial issues.

Following the FAIR recommendation (https://zenodo.org/record/1285272) to establish an *Open Data Mandate* for (publicly funded) research, it is a crucial prerequisite to enable secondary use of 'omics' data to a broader extent. Only if data are technically accessible by using open source file formats including the relevant data and information (e.g. relevant metadata), the consistent implementation of Big Data approaches is possible.

There are current substantial efforts towards more *open file formats*. For example, the HUPO-PSI (5.5) has made progress in this regard.

However, in order to establish a future *open data culture* there is a need for an improved use of *open source standard file formats* which should be vendor independent / non-proprietary.

Linking large numbers of heterogeneous data sources, e.g. coupling proteomics databases to different sources of electronic health records remains challenging.

Robust and feasible approaches have to be established in order to link different sources of 'omics' data with different sources of clinical and non-clinical data.

The used data (file) formats should be harmonised to the best, feasible extent. In order to establish an *Open Data Mandate,* it is crucial to identify or develop *open source file formats* which include the relevant data and information (e.g. relevant metadata). Regulatory agencies may contribute by providing advice which data (file) formats and / or attributes of data formats are acceptable for regulatory purpose.

Guidance should be provided which *data (file) formats* are acceptable for which regulatory use.

8.5. Strengthening the development and harmonisation of data standards

It is important to define the requirements or standards to which databases should adhere. It is encouraged to limit the use of *data standards* to the feasible minimum.

There are general efforts and approaches as lined out in the FAIR recommendations or Critical Path Initiative (CPI) to harmonise the information provided via *data standards*.

Suitable data standards are required to allow an appropriate integration of large scale (Big) Data sets. Suitable and appropriate *data standards* like CDISC standards should be identified and -in case this is necessary- adapted for the use in Big Data approaches, for instance, by establishing linkage and including specific common data elements.

A main question is the validity/accuracy of the data; thus, quality attributes should be included in order to allow appropriate selection, analysis and interpretation of data sets.

Data standards should be platform-independent, appropriately validated and freely available. However, even if harmonisation should be supported to the best extent, the complexity, particularly in the field of other 'omics' technologies, limits the feasibility to harmonise data standards rigorously.

It is encouraged to limit the number of used *data standards* to the feasible minimum.

Suitable and appropriate *data standards* like CDISC standards should be identified and -in case this is necessary- adapted for the use in Big Data approaches, for instance, by establishing linkage and including specific common data elements. Quality attributes should be included in order to allow appropriate selection, analysis and interpretation of data sets. *Data standards* should be platform-independent, appropriately validated and freely available.

Guidance should be provided which data standards are acceptable for which regulatory purpose, depending on the impact on diagnostic, medical or regulatory decisions.

8.6. Regulatory recognition of clinical relevance and prognostic / predictive value

The general procedures and relevant aspects (such as the definition of the Context(s) of Use) for the establishment and regulatory recognition of clinical relevance and prognostic value are specified in *Qualification of novel methodologies for drug development* EMA/CHMP/SAWP/72894/2008**Error! Bookmark not defined.**; for further explanation please refer to the section 5.9.

However, appropriate regulatory recognition of Big Data *'Bioanalytical omics'* approaches will also challenge regulatory systems, e.g. addressing comprehensively, but feasible particular aspects of the endpoint selection (in case the mechanistic understanding underlying mechanisms and pathways is limited). If the direct assessment of the standard of truth is not possible (e.g. practical / clinical reasons or lack of definition of conditions intended to be diagnosed), there is the need to establish robust and meaningful 'surrogate standard of truth' – accepted by the regulatory and health care systems.

It will be in the remit of the regulator network in Europe (and globally) to specify the type of evidence expected for each potential use of Big Data ('omics') approaches.

This includes general advice provided via relevant guidelines and particular guidance provided for particular development projects via scientific advice and qualification advice.

It will be in the remit of the regulator network in Europe (and globally) to specify the type of evidence expected for each potential use of Big Data ('omics') approaches. This includes general advice provided via relevant guidelines and particular guidance provided for particular development projects via scientific advice and qualification advice.

The general procedures and context for the regulatory recognition of results derived from of Big Data 'omics' approaches are provided by *Qualification of novel methodologies for drug development* established within the European regulatory network.

Specific and targeted adaptations may be required for general questions (feeding in the development of relevant guidelines) or specific development projects. These adaptations should be specified by the regulatory agencies based on the particular methodical requirements and on practical experiences with the application of these new technologies.

8.7. Bioinformatics and statistical considerations

The methods of data analysis used for Big Data 'omics' approaches (i.e. models and algorithms) are subject to the growing field of data science which combines methods from various disciplines such as biostatistics, mathematical modelling and simulation, bio-informatics and computer science including data-integration/machine learning and high-performance computing.

The workflow of Big Data analysis approaches can be generally divided in two steps:

Big Data analyses are typically preceded ***Data Processing*** to integrate data from different sources followed by ***Data Analysis & Interpretation***: Based on these pre-processed (and curated) data, inferences can be drawn using methods from biostatistics, mathematical modelling and simulation, bioinformatics and machine learning. While data processing can largely be automated, model and method development, validation and interpretation strongly rely on expert knowledge of data scientists.

Due to the requirements for assessing appropriateness of the regulatory use of Big Data processing and analysis/interpretation the adaptation of existing and the establishment of new regulatory approaches will be required. There is the challenge for the regulatory system to ensure the reliability of results, but also to support and encourage the development and establishment of new methodologies in order to strengthen and improve the development of innovative treatments. The "containerisation" of workflows holds great promise regarding reproducibility of data processing and data analysis pipelines. Case studies would be helpful to define requirements and to adapt or establish related regulatory workflows.

Big Data 'omics' approaches rely on innovative data science driven analytical tools (i.e. models and algorithms). To assess the appropriateness of the regulatory use of Big Data processing and analysis/interpretation, the adaptation of existing and/or the establishment of new regulatory approaches will be required in order to confirm diagnostic performance and clinical utility.

There is the need of specifying (novel) requirements defining which approaches are considered as confirmatory. It will be in the remit of regulatory agencies to define methodical requirements for data analytics, depending on the impact on diagnostic, medical or regulatory decisions.

8.8. Knowledge /expertise gaps within the European regulatory network

In reference to the issues discussed above, specialised expertise and experience in the field of data science which combines methods from various disciplines such as biostatistics, mathematical modelling and simulation, bioinformatics and computer science is required to address regulatory needs and questions in the future appropriately. Even, if highly specialised knowledge will only be required in certain cases, comprehensive and broad expertise will be required as in-house knowledge e.g. for an appropriate assessment of a marketing authorisation application. As different disciplines and capabilities are required to ensure the operational capacity of the regulatory system, the relevant capacities and resources should be implemented and cross-linked in a coordinated manner.

The use of Big Data 'omics' approaches and their analysis call for new regulatory strategies and guidance to achieve their full beneficial potential. In order to ensure appropriate regulatory assessment of relevant submissions and to strengthen development capabilities, the expertise of data scientists from various disciplines (e.g. mathematical modelling and simulation, bioinformatics and computer science) will be needed. The required capacities should be built up in a collaborative effort to ensure efficient implementation at the European level. The result should be a regulatory network in Europe which is capable to address related challenges appropriately, but also to support innovation to the best extent, to ensure optimal results for European societies and patients.

9. Recommendations

Topic	Core Recommendation	Reinforcing Actions	Metric/ KPI
Comprehensiveness of available data sets [bioanalytical omics]	Implement an analysis to assess the completeness of available data and the potential impact of missing data/ information.	• Establish a suitable framework, specifying the conduct of 'omics' Big Data analysis approaches.	Increase the number of validated/qualified 'omics' Big Data biomarkers.
Data Quality [samples and documentation]	Guidance should be provided on acceptability on Big Data sets to support regulatory decision making.	• Quality attributes of (Big)' data sets need to be defined by regulators including appropriate data (file) formats and data standards. • Quality attributes should be included in order to allow appropriate selection, analysis and interpretation of data sets.	
Bioanalytical method validation	Clear guidance should be provided for the bioanalytical method validation suitable for the complexity of bioanalytical omics techniques.	• Standards for method validation should be specified by / or in close collaboration with relevant competent authorities. • Quality relevant aspects of bioanalytical method validation as well as Data Processing and Data Analysis & Interpretation should be addressed in the specific recommendations.	

Supporting the harmonisation and sharing of data (file) formats	Data (file) formats should be harmonised as much as possible.	• In order to establish an *Open Data Mandate* it is crucial to identify or develop *open source file formats* which include the relevant data and information (e.g. relevant metadata). • Regulatory agencies should contribute by providing advice on which *data file formats* and / or attributes of data formats are acceptable for regulatory purpose.	Increase the number of available, relevant and harmonised 'omics' Big Data data sets acceptable for regulatory decision making.
Strengthening the development and harmonisation of data standards	It is encouraged to limit the number of used *data standards* to the minimum.	• Suitable and appropriate *data standards* e.g. CDISC should be identified and if necessary, adapted for the use in Big Data approaches. • *Data standards* should be platform-independent, appropriately validated and freely available.	
Regulatory recognition of clinical relevance and prognostic/ predictive value of omics	Regulatory agencies should clearly articulate what degree of evidence is acceptable in order to support regulatory decision making, highlighting their value as prognostic markers.	• In line with guidelines developed for genomics and for novel modelling approaches (PBPK), similar guidelines need to be developed for other omics. • Regulatory guidance/advice should be provided via qualification of novel methodologies for medicine development.	The aim is to establish a framework of relevant guidance documents for regulatory use of Big Data ('omics') approaches. This should be accompanied by targeted and qualified scientific advice for particular projects and scientific questions.
Bioinformatics and statistical considerations	Specific guidance should be provided on the bioinformatic and statistical requirements for regulatory acceptability of Big Data analyses.	• The adaptation of existing or the establishment of new regulatory approaches might be required in order to confirm diagnostic performance and clinical utility of 'omics' Big Data strategies.	
Knowledge /expertise gaps within the European regulatory network	To ensure appropriate assessment of regulatory submissions expertise in various disciplines (e.g. mathematical modelling and simulation, bioinformatics and computer sciences) will be needed.	• Recruitment of appropriate expertise where none exists in the regulatory network. • The required capacities should be trained through a focused training programme on a European level. • Case studies would be an efficient tool to train and strengthen the capacities of the European regulatory network in the fields of computer science including data-integration/machine learning and high-performance computing.	Increase the number of competent assessors.

10. References

Aebersold, Ruedi; Mann, Matthias (2016): Mass-spectrometric exploration of proteome structure and function. In: *Nature* 537 (7620), S. 347–355. DOI: 10.1038/nature19949.

Anderson, N. Leigh (2010): The clinical plasma proteome. A survey of clinical assays for proteins in plasma and serum. In: *Clinical chemistry* 56 (2), S. 177–185. DOI: 10.1373/clinchem.2009.126706.

Balgley, Brian M.; Wang, Weijie; Song, Tao; Fang, Xueping; Yang, Li; Lee, Cheng S. (2008): Evaluation of confidence and reproducibility in quantitative proteomics performed by a capillary

isoelectric focusing-based proteomic platform coupled with a spectral counting approach. In: *Electrophoresis* 29 (14), S. 3047–3054. DOI: 10.1002/elps.200800050.

Bell, Alexander W.; Deutsch, Eric W.; Au, Catherine E.; Kearney, Robert E.; Beavis, Ron; Sechi, Salvatore et al. (2009): A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. In: *Nature methods* 6 (6), S. 423–430. DOI: 10.1038/nmeth.1333.

Boyanova, Desislava; Nilla, Santosh; Klau, Gunnar W.; Dandekar, Thomas; Müller, Tobias; Dittrich, Marcus (2014): Functional module search in protein networks based on semantic similarity improves the analysis of proteomics data. In: *Molecular & cellular proteomics : MCP* 13 (7), S. 1877–1889. DOI: 10.1074/mcp.M113.032839.

Britten, Cedrik M.; Singh-Jasuja, Harpreet; Flamion, Bruno; Hoos, Axel; Huber, Christoph; Kallen, Karl-Josef et al. (2013): The regulatory landscape for actively personalized cancer immunotherapies. In: *Nature biotechnology* 31 (10), S. 880–882. DOI: 10.1038/nbt.2708.

Budin-Ljøsne, Isabelle; Burton, Paul; Isaeva, Julia; Gaye, Amadou; Turner, Andrew; Murtagh, Madeleine J. et al. (2015): DataSHIELD. An ethically robust solution to multiple-site individual-level data analysis. In: *Public health genomics* 18 (2), S. 87–96. DOI: 10.1159/000368959.

Chambers, Matthew C.; MacLean, Brendan; Burke, Robert; Amodei, Dario; Ruderman, Daniel L.; Neumann, Steffen et al. (2012): A cross-platform toolkit for mass spectrometry and proteomics. In: *Nature biotechnology* 30 (10), S. 918–920. DOI: 10.1038/nbt.2377.

Chamrad, Daniel; Meyer, Helmut E. (2005): Valid data from large-scale proteomics studies. In: *Nature methods* 2 (9), S. 647–648. DOI: 10.1038/nmeth0905-647.

Chandramouli, Kondethimmanahalli; Qian, Pei-Yuan (2009): Proteomics. Challenges, techniques and possibilities to overcome biological sample complexity. In: *Human genomics and proteomics : HGP* 2009. DOI: 10.4061/2009/239204.

Côté, Richard G.; Reisinger, Florian; Martens, Lennart (2010): jmzML, an open-source Java API for mzML, the PSI standard for MS data. In: *Proteomics* 10 (7), S. 1332–1335. DOI: 10.1002/pmic.200900719.

Delmotte, Nathanaël; Lasaosa, Maria; Tholey, Andreas; Heinzle, Elmar; van Dorsselaer, Alain; Huber, Christian G. (2009): Repeatability of peptide identifications in shotgun proteome analysis employing off-line two-dimensional chromatographic separations and ion-trap MS. In: *Journal of separation science* 32 (8), S. 1156–1164. DOI: 10.1002/jssc.200800615.

Deutsch, Eric (2008): mzML. A single, unifying data format for mass spectrometer output. In: *Proteomics* 8 (14), S. 2776–2777. DOI: 10.1002/pmic.200890049.

Deutsch, Eric W.; Csordas, Attila; Sun, Zhi; Jarnuczak, Andrew; Perez-Riverol, Yasset; Ternent, Tobias et al. (2017): The ProteomeXchange consortium in 2017. Supporting the cultural change in proteomics public data deposition. In: *Nucleic acids research* 45 (D1), D1100-D1106. DOI: 10.1093/nar/gkw936.

Diggs, Laurence P.; Hsueh, Eddy C. (2017): Utility of PD-L1 immunohistochemistry assays for predicting PD-1/PD-L1 inhibitor response. In: *Biomarker research* 5, S. 12. DOI: 10.1186/s40364-017-0093-8.

Domon, Bruno; Aebersold, Ruedi (2010): Options and considerations when selecting a quantitative proteomics strategy. In: *Nature biotechnology* 28 (7), S. 710–721. DOI: 10.1038/nbt.1661.

Elias, Joshua E.; Haas, Wilhelm; Faherty, Brendan K.; Gygi, Steven P. (2005): Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. In: *Nature methods* 2 (9), S. 667–675. DOI: 10.1038/nmeth785.

Everett, Jeremy R.; Loo, Ruey Leng; Pullen, Francis S. (2013): Pharmacometabonomics and personalized medicine. In: *Annals of clinical biochemistry* 50 (Pt 6), S. 523–545. DOI: 10.1177/0004563213497929.

Farrah, Terry; Deutsch, Eric W.; Omenn, Gilbert S.; Sun, Zhi; Watts, Julian D.; Yamamoto, Tadashi et al. (2014): State of the human proteome in 2013 as viewed through PeptideAtlas. Comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. In: *Journal of proteome research* 13 (1), S. 60–75. DOI: 10.1021/pr4010037.

Farriol-Mathis, Nathalie; Garavelli, John S.; Boeckmann, Brigitte; Duvaud, Séverine; Gasteiger, Elisabeth; Gateau, Alain et al. (2004): Annotation of post-translational modifications in the Swiss-Prot knowledge base. In: *Proteomics* 4 (6), S. 1537–1550. DOI: 10.1002/pmic.200300764.

Gloriam, David E.; Orchard, Sandra; Bertinetti, Daniela; Björling, Erik; Bongcam-Rudloff, Erik; Borrebaeck, Carl A. K. et al. (2010): A community standard format for the representation of protein affinity reagents. In: *Molecular & cellular proteomics : MCP* 9 (1), S. 1–10. DOI: 10.1074/mcp.M900185-MCP200.

Grant, Russell P.; Hoofnagle, Andrew N. (2014): From lost in translation to paradise found. Enabling protein biomarker method transfer by mass spectrometry. In: *Clinical chemistry* 60 (7), S. 941–944. DOI: 10.1373/clinchem.2014.224840.

Gu, Qiang; Yu, Li-Rong (2014): Proteomics quality and standard. From a regulatory perspective. In: *Journal of proteomics* 96, S. 353–359. DOI: 10.1016/j.jprot.2013.11.024.

Holmes, Elaine; Wijeyesekera, Anisha; Taylor-Robinson, Simon D.; Nicholson, Jeremy K. (2015): The promise of metabolic phenotyping in gastroenterology and hepatology. In: *Nature reviews. Gastroenterology & hepatology* 12 (8), S. 458–471. DOI: 10.1038/nrgastro.2015.114.

Ivanov, Alexander R.; Colangelo, Christopher M.; Dufresne, Craig P.; Friedman, David B.; Lilley, Kathryn S.; Mechtler, Karl et al. (2013): Interlaboratory studies and initiatives developing standards for proteomics. In: *Proteomics* 13 (6), S. 904–909. DOI: 10.1002/pmic.201200532.

Kapp, Eugene A.; Schütz, Frédéric; Connolly, Lisa M.; Chakel, John A.; Meza, Jose E.; Miller, Christine A. et al. (2005): An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms. Sensitivity and specificity analysis. In: *Proteomics* 5 (13), S. 3475–3490. DOI: 10.1002/pmic.200500126.

Keller, Andrew; Eng, Jimmy; Zhang, Ning; Li, Xiao-jun; Aebersold, Ruedi (2005): A uniform proteomics MS/MS analysis platform utilizing open XML file formats. In: *Molecular systems biology* 1, 2005.0017. DOI: 10.1038/msb4100024.

Kessner, Darren; Chambers, Matt; Burke, Robert; Agus, David; Mallick, Parag (2008): ProteoWizard. Open source software for rapid proteomics tools development. In: *Bioinformatics (Oxford, England)* 24 (21), S. 2534–2536. DOI: 10.1093/bioinformatics/btn323.

Khoury, George A.; Baliban, Richard C.; Floudas, Christodoulos A. (2011): Proteome-wide post-translational modification statistics. Frequency analysis and curation of the swiss-prot database. In: *Scientific reports* 1. DOI: 10.1038/srep00090.

Kislinger, Thomas; Gramolini, Anthony O.; MacLennan, David H.; Emili, Andrew (2005): Multidimensional protein identification technology (MudPIT). Technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue. In: *Journal of the American Society for Mass Spectrometry* 16 (8), S. 1207–1220. DOI: 10.1016/j.jasms.2005.02.015.

Kowalewski, Daniel J.; Schuster, Heiko; Backert, Linus; Berlin, Claudia; Kahn, Stefan; Kanz, Lothar et al. (2015): HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (2), E166-75. DOI: 10.1073/pnas.1416389112.

Li, Bo; Liao, Bo (2017): Protein Complexes Prediction Method Based on Core-Attachment Structure and Functional Annotations. In: *International journal of molecular sciences* 18 (9). DOI: 10.3390/ijms18091910.

Lindon, John C.; Nicholson, Jeremy K.; Holmes, Elaine; Keun, Hector C.; Craig, Andrew; Pearce, Jake T. M. et al. (2005): Summary recommendations for standardization and reporting of metabolic analyses. In: *Nature biotechnology* 23 (7), S. 833–838. DOI: 10.1038/nbt0705-833.

Liu, Hongbin; Sadygov, Rovshan G.; Yates, John R. (2004): A model for random sampling and estimation of relative protein abundance in shotgun proteomics. In: *Analytical chemistry* 76 (14), S. 4193–4201. DOI: 10.1021/ac0498563.

Mann, Matthias (2009): Comparative analysis to guide quality improvements in proteomics. In: *Nature methods* 6 (10), S. 717–719.

Martens, Lennart; Chambers, Matthew; Sturm, Marc; Kessner, Darren; Levander, Fredrik; Shofstahl, Jim et al. (2011): mzML--a community standard for mass spectrometry data. In: *Molecular & cellular proteomics : MCP* 10 (1), R110.000133. DOI: 10.1074/mcp.R110.000133.

Martens, Lennart; Vizcaíno, Juan Antonio (2017): A Golden Age for Working with Public Proteomics Data. In: *Trends in biochemical sciences* 42 (5), S. 333–341. DOI: 10.1016/j.tibs.2017.01.001.

Mertins, Philipp; Mani, D. R.; Ruggles, Kelly V.; Gillette, Michael A.; Clauser, Karl R.; Wang, Pei et al. (2016): Proteogenomics connects somatic mutations to signalling in breast cancer. In: *Nature* 534 (7605), S. 55–62. DOI: 10.1038/nature18003.

Mihai, Simona; Codrici, Elena; Popescu, Ionela Daniela; Enciu, Ana-Maria; Rusu, Elena; Zilisteanu, Diana et al. (2016): Proteomic Biomarkers Panel. New Insights in Chronic Kidney Disease. In: *Disease markers* 2016, S. 3185232. DOI: 10.1155/2016/3185232.

Nicholson, Jeremy K.; Holmes, Elaine; Kinross, James M.; Darzi, Ara W.; Takats, Zoltan; Lindon, John C. (2012): Metabolic phenotyping in clinical and surgical environments. In: *Nature* 491 (7424), S. 384–392. DOI: 10.1038/nature11708.

O'Donnell, Valerie B.; Murphy, Robert C.; Watson, Steve P. (2014): Platelet lipidomics. Modern day perspective on lipid discovery and characterization in platelets. In: *Circulation research* 114 (7), S. 1185–1203. DOI: 10.1161/CIRCRESAHA.114.301597.

Omenn, Gilbert S.; States, David J.; Adamski, Marcin; Blackwell, Thomas W.; Menon, Rajasree; Hermjakob, Henning et al. (2005): Overview of the HUPO Plasma Proteome Project. Results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. In: *Proteomics* 5 (13), S. 3226–3245. DOI: 10.1002/pmic.200500358.

Paulovich, Amanda G.; Whiteaker, Jeffrey R. (2016): Quantifying the human proteome. In: *Nature biotechnology* 34 (10), S. 1033–1034. DOI: 10.1038/nbt.3695.

Pedrioli, Patrick G. A.; Eng, Jimmy K.; Hubley, Robert; Vogelzang, Mathijs; Deutsch, Eric W.; Raught, Brian et al. (2004): A common open representation of mass spectrometry data and its application to proteomics research. In: *Nature biotechnology* 22 (11), S. 1459–1466. DOI: 10.1038/nbt1031.

Peng, Junmin; Elias, Joshua E.; Thoreen, Carson C.; Licklider, Larry J.; Gygi, Steven P. (2003): Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis. The yeast proteome. In: *Journal of proteome research* 2 (1), S. 43–50.

Perez-Riverol, Yasset; Alpi, Emanuele; Wang, Rui; Hermjakob, Henning; Vizcaíno, Juan Antonio (2015): Making proteomics data accessible and reusable. Current state of proteomics databases and repositories. In: *Proteomics* 15 (5-6), S. 930–949. DOI: 10.1002/pmic.201400302.

Ponomarenko, Elena A.; Poverennaya, Ekaterina V.; Ilgisonis, Ekaterina V.; Pyatnitskiy, Mikhail A.; Kopylov, Arthur T.; Zgoda, Victor G. et al. (2016): The Size of the Human Proteome. The Width and Depth. In: *International journal of analytical chemistry* 2016, S. 7436849. DOI: 10.1155/2016/7436849.

Rammensee, Hans-Georg; Singh-Jasuja, Harpreet (2013): HLA ligandome tumor antigen discovery for personalized vaccine approach. In: *Expert review of vaccines* 12 (10), S. 1211–1217. DOI: 10.1586/14760584.2013.836911.

Resing, Katheryn A.; Meyer-Arendt, Karen; Mendoza, Alex M.; Aveline-Wolf, Lauren D.; Jonscher, Karen R.; Pierce, Kevin G. et al. (2004): Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. In: *Analytical chemistry* 76 (13), S. 3556–3568. DOI: 10.1021/ac035229m.

Schaab, Christoph; Geiger, Tamar; Stoehr, Gabriele; Cox, Juergen; Mann, Matthias (2012): Analysis of high accuracy, quantitative proteomics data in the MaxQB database. In: *Molecular & cellular proteomics : MCP* 11 (3), M111.014068. DOI: 10.1074/mcp.M111.014068.

Schramm, Thorsten; Hester, Alfons; Klinkert, Ivo; Both, Jean-Pierre; Heeren, Ron M. A.; Brunelle, Alain et al. (2012): imzML--a common data format for the flexible exchange and processing of mass spectrometry imaging data. In: *Journal of proteomics* 75 (16), S. 5106–5110. DOI: 10.1016/j.jprot.2012.07.026.

Searle, Brian C. (2010): Scaffold. A bioinformatic tool for validating MS/MS-based proteomic studies. In: *Proteomics* 10 (6), S. 1265–1269. DOI: 10.1002/pmic.200900437.

Sethi, Sumit; Brietzke, Elisa (2017): Recent advances in lipidomics. Analytical and clinical perspectives. In: *Prostaglandins & other lipid mediators* 128-129, S. 8–16. DOI: 10.1016/j.prostaglandins.2016.12.002.

Slebos, Robbert J. C.; Brock, Jonathan W. C.; Winters, Nancy F.; Stuart, Sarah R.; Martinez, Misti A.; Li, Ming et al. (2008): Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. In: *Journal of proteome research* 7 (12), S. 5286–5294. DOI: 10.1021/pr8004666.

Smith, Lloyd M.; Kelleher, Neil L. (2013): Proteoform. A single term describing protein complexity. In: *Nature methods* 10 (3), S. 186–187. DOI: 10.1038/nmeth.2369.

Steffen, Pascal; Kwiatkowski, Marcel; Robertson, Wesley D.; Zarrine-Afsar, Arash; Deterra, Diana; Richter, Verena; Schlüter, Hartmut (2016): Protein species as diagnostic markers. In: *Journal of proteomics* 134, S. 5–18. DOI: 10.1016/j.jprot.2015.12.015.

Stoevesandt, Oda; Taussig, Michael J. (2012): European and international collaboration in affinity proteomics. In: *New biotechnology* 29 (5), S. 511–514. DOI: 10.1016/j.nbt.2012.05.003.

Sturm, Marc; Bertsch, Andreas; Gröpl, Clemens; Hildebrandt, Andreas; Hussong, Rene; Lange, Eva et al. (2008): OpenMS - an open-source software framework for mass spectrometry. In: *BMC bioinformatics* 9, S. 163. DOI: 10.1186/1471-2105-9-163.

Tabb, David L. (2013): Quality assessment for clinical proteomics. In: *Clinical biochemistry* 46 (6), S. 411–420. DOI: 10.1016/j.clinbiochem.2012.12.003.

Tabb, David L.; Vega-Montoto, Lorenzo; Rudnick, Paul A.; Variyath, Asokan Mulayath; Ham, Amy-Joan L.; Bunk, David M. et al. (2010): Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. In: *Journal of proteome research* 9 (2), S. 761–776. DOI: 10.1021/pr9006365.

Taussig, Michael J.; Stoevesandt, Oda; Borrebaeck, Carl A. K.; Bradbury, Andrew R.; Cahill, Dolores; Cambillau, Christian et al. (2007): ProteomeBinders. Planning a European resource of affinity reagents for analysis of the human proteome. In: *Nature methods* 4 (1), S. 13–17. DOI: 10.1038/nmeth0107-13.

Taylor, Chris F.; Field, Dawn; Sansone, Susanna-Assunta; Aerts, Jan; Apweiler, Rolf; Ashburner, Michael et al. (2008): Promoting coherent minimum reporting guidelines for biological and biomedical investigations. The MIBBI project. In: *Nature biotechnology* 26 (8), S. 889–896. DOI: 10.1038/nbt.1411.

Taylor, Chris F.; Paton, Norman W.; Lilley, Kathryn S.; Binz, Pierre-Alain; Julian, Randall K.; Jones, Andrew R. et al. (2007): The minimum information about a proteomics experiment (MIAPE). In: *Nature biotechnology* 25 (8), S. 887–893. DOI: 10.1038/nbt1329.

Triebl, Alexander; Hartler, Jürgen; Trötzmüller, Martin; C Köfeler, Harald (2017): Lipidomics. Prospects from a technological perspective. In: *Biochimica et biophysica acta* 1862 (8), S. 740–746. DOI: 10.1016/j.bbalip.2017.03.004.

Tripathi, Shashank; Pohl, Marie O.; Zhou, Yingyao; Rodriguez-Frandsen, Ariel; Wang, Guojun; Stein, David A. et al. (2015): Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. In: *Cell host & microbe* 18 (6), S. 723–735. DOI: 10.1016/j.chom.2015.11.002.

Uhlen, Mathias; Zhang, Cheng; Lee, Sunjae; Sjöstedt, Evelina; Fagerberg, Linn; Bidkhori, Gholamreza et al. (2017): A pathology atlas of the human cancer transcriptome. In: *Science (New York, N.Y.)* 357 (6352). DOI: 10.1126/science.aan2507.

van Midwoud, Paul M.; Rieux, Laurent; Bischoff, Rainer; Verpoorte, Elisabeth; Niederländer, Harm A. G. (2007): Improvement of recovery and repeatability in liquid chromatography-mass spectrometry analysis of peptides. In: *Journal of proteome research* 6 (2), S. 781–791. DOI: 10.1021/pr0604099.

Varjosalo, Markku; Sacco, Roberto; Stukalov, Alexey; van Drogen, Audrey; Planyavsky, Melanie; Hauri, Simon et al. (2013): Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. In: *Nature methods* 10 (4), S. 307–314. DOI: 10.1038/nmeth.2400.

Vialas, Vital; Colomé-Calls, Núria; Abian, Joaquín; Aloria, Kerman; Alvarez-Llamas, Gloria; Antúnez, Oreto et al. (2017): A multicentric study to evaluate the use of relative retention times in targeted proteomics. In: *Journal of proteomics* 152, S. 138–149. DOI: 10.1016/j.jprot.2016.10.014.

Vidova, Veronika; Spacil, Zdenek (2017): A review on mass spectrometry-based quantitative proteomics. Targeted and data independent acquisition. In: *Analytica chimica acta* 964, S. 7–23. DOI: 10.1016/j.aca.2017.01.059.

Vihervaara, Terhi; Suoniemi, Matti; Laaksonen, Reijo (2014): Lipidomics in drug discovery. In: *Drug discovery today* 19 (2), S. 164–170. DOI: 10.1016/j.drudis.2013.09.008.

Wang, Xia (2017): Statistical Assessment of QC Metrics on Raw LC-MS/MS Data. In: *Methods in molecular biology (Clifton, N.J.)* 1550, S. 325–337. DOI: 10.1007/978-1-4939-6747-6_22.

Washburn, M. P.; Wolters, D.; Yates, J. R. (2001): Large-scale analysis of the yeast proteome by multidimensional protein identification technology. In: *Nature biotechnology* 19 (3), S. 242–247. DOI: 10.1038/85686.

Washburn, Michael P.; Ulaszek, Ryan R.; Yates, John R. (2003): Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology. In: *Analytical chemistry* 75 (19), S. 5054–5061.

Whiteaker, Jeffrey R.; Halusa, Goran N.; Hoofnagle, Andrew N.; Sharma, Vagisha; MacLean, Brendan; Yan, Ping et al. (2016): Using the CPTAC Assay Portal to Identify and Implement Highly Characterized Targeted Proteomics Assays. In: *Methods in molecular biology (Clifton, N.J.)* 1410, S. 223–236. DOI: 10.1007/978-1-4939-3524-6_13.

Wilkins, Marc (2009): Proteomics data mining. In: *Expert review of proteomics* 6 (6), S. 599–603. DOI: 10.1586/epr.09.81.

Yang, Kui; Han, Xianlin (2016): Lipidomics. Techniques, Applications, and Outcomes Related to Biomedical Sciences. In: *Trends in biochemical sciences* 41 (11), S. 954–969. DOI: 10.1016/j.tibs.2016.08.010.

Zhang, Bo; Käll, Lukas; Zubarev, Roman A. (2016): DeMix-Q. Quantification-Centered Data Processing Workflow. In: *Molecular & cellular proteomics : MCP* 15 (4), S. 1467–1478. DOI: 10.1074/mcp.O115.055475.