Dissertation zur Erlangung des Doktorgrades

der Fakultät für Chemie und Pharmazie

der Ludwig-Maximilians-Universität München

Advancing Computational Methods for Mass Spectrometry-Based Proteomics, Metabolomics, and Analysis of Multi-Omics Datasets

Hamid Hamzeiy

aus

Tabriz, Iran

2021

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Jürgen Cox betreut
und von Herrn Prof. Dr. Christoph W. Turck
von der Fakultät für Chemie und Pharmazie vertreten.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 13.01.2021

Hamid Hamzeiy

.....

Dissertation eingereicht am	13.01.2021
1. Gutachterin / 1. Gutachter:	Prof. Dr. Christoph W. Turck
2. Gutachterin / 2. Gutachter:	Dr. Jürgen Cox
Mündliche Prüfung am	11.02.2021

To being perseverant…

Summary

Undoubtedly, the current century is witness to an unprecedented speed in advancements within biological sciences, which are owed to the immense technological progress in the analytical tools and methods utilized, and to the dawn of the fast developing fields of omics and bioinformatics. Omics allows the collection of holistic data on several different biomolecule classes, and bioinformatics makes it possible to explore and understand the vast amounts of data produced. The most mature omics fields, in terms of both hardware and software, are genomics and transcriptomics, based on next generation sequencing (NGS) technologies. With the introduction of electrospray ionization and high-resolution mass spectrometry, liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS), has made significant leaps for the fields of metabolomics and proteomics.

One promising method for LC-MS/MS-based proteomics is data independent acquisition (DIA), which requires advanced data analysis algorithms. MaxDIA, within the MaxQuant software for the processing of LC-MS/MS-based proteomics data, is introduced here. It comes with an accurate false discovery rate estimation of the peptide and protein identification based on measured and predicted spectrum libraries. When compared to the state of the art, MaxDIA also delivers comprehensive proteome coverages and lower coefficients of variation in protein quantification.

Bioinformatics tools for the analysis of metabolomics data generally follow the same principles and steps as proteomics software, but due the huge numbers of metabolites and immense complexity of metabolomics data, much work is still needed to bring metabolomics software to the level of maturity of their proteomics equivalents. MaxQuant is a time tested and widely accepted software for the processing of proteomics data, which was first recognized for its cutting-edge nonlinear recalibration for reaching superior precursor mass accuracy, which helps significantly improve peptide identifications. Here, following this direction, a new algorithm within MaxQuant for improving mass accuracy in metabolomics data is introduced, which utilizes a novel metabolite library-based mass recalibration algorithm.

The many types of omics data available today present a great opportunity for developing approaches to combine such data in order to infer new knowledge, often termed multi-omics studies. A robust approach to this end is to utilize prior knowledge on the relationships of the various major biomolecules in question, which are often depicted in network structures where the nodes of the network depict biomolecules and the edges correspond to an interaction. To implement this approach, Metis is introduced, a new

plugin for the Perseus software aimed at analyzing quantitative multi-omics data based on metabolic pathways.

This thesis includes four publications, the first of which is a review article on computational metabolomics as a part of the introduction, listed below:

1. **Hamzeiy, Hamid**, and Jürgen Cox. 2017. "What Computational Non-Targeted Mass Spectrometry-Based Metabolomics Can Gain from Shotgun Proteomics." Current Opinion in Biotechnology 43: 141–46. https://doi.org/10.1016/j.copbio.2016.11.014.

2. Sinitcyn, Pavel, Shivani Tiwary, Jan Rudolph, Petra Gutenbrunner, Christoph Wichmann, Şule Yllmaz, **Hamid Hamzeiy**, Favio Salinas, and Jürgen Cox. 2018. "MaxQuant Goes Linux." Nature Methods 15 (6): 401. https://doi.org/10.1038/s41592-018-0018-y.

3. Pavel Sinitcyn, **Hamid Hamzeiy**, Favio Salinas Soto, Daniel Itzhak, Frank McCarthy, Christoph Wichmann, Martin Steger, Uli Ohmayer, Ute Distler, Stephanie Kaspar-Schoenefeld, Nikita Prianichnikov, Şule Yılmaz, Jan Daniel Rudolph, Stefan Tenzer, Yasset Perez-Riverol, Nagarjuna Nagaraj, Sean J. Humphrey and Jürgen Cox. "MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics." Submitted to Nature Biotechnology, 2020

4. **Hamid Hamzeiy**, Daniela Ferretti, Maria S. Robles, and Jürgen Cox. "Perseus plugin 'Metis' for metabolic pathway-centered quantitative multi-omics data analysis supporting static and time-series experimental designs." Submitted to Cell Systems, 2021

Contents

1. Introduction

Biological sciences in the current century are becoming an ever more data-driven endeavor (MacKlin, 2019). This is both due to the shear increase in the availability of vast resources of high-throughput omics data, and also improvements in methods and algorithms used for the generation, processing and analysis of such data (Chavan, Shaughnessy and Edmondson, 2011; Cox and Mann, 2011; Sinitcyn, Rudolph and Cox, 2018). One can say that biological sciences have now truly entered the fourth paradigm of science in efforts for the exploration and understanding of biological systems (Figure 1.1). This has brought with it a new push to advance a dynamic field of research within biology, namely bioinformatics, which has emerged to be of central importance in many aspects of novel experimental design, and knowledge discovery (Gauthier *et al.*, 2019). Bioinformatics efforts aim to bridge the gap between biology and informatics and work to enable biological researchers to effectively analyze the data gathered to reach a deeper insight into various aspects of living systems (Luscombe, Greenbaum and Gerstein, 2001).

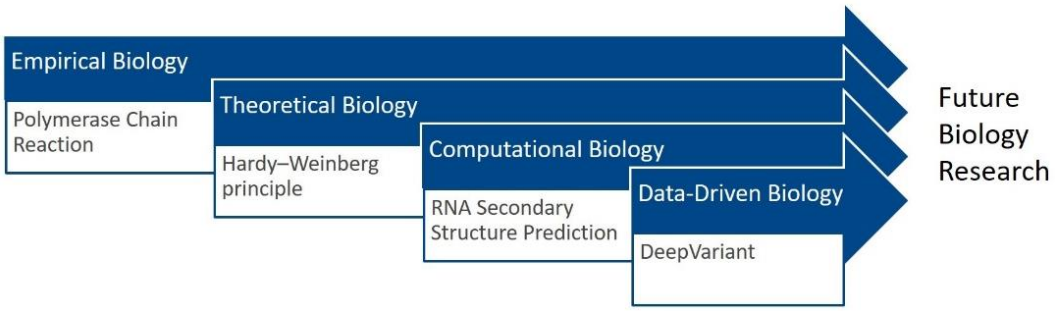


Figure 1.1: Paradigms in biology (adapted from (Agrawal and Choudhary, 2016)*). Each arrow depicts a paradigm in biology, with an example for a development in that direction.*

Current bioinformatics solutions have been essential in propelling our understanding of many aspects of biology, but there still exists a great deal that remains to be developed and thus, it is ever more important to focus efforts on novel tools and algorithms that can handle such large quantities of highly complex data (Fuller *et al.*, 2013; Gauthier *et al.*, 2019). In this introductory chapter, omics are discussed in general, along with computational proteomics, computational metabolomics, and methods for multi-omics data analysis.

1.1 Omics

The capacity to holistically collect data, and study any living entity, ranging from single cells to large multicellular organisms, has given way to new fields of study, commonly labelled as omics (Karahalil, 2016). Naturally, the generation of large omics datasets has brought with it, both a novel set of opportunities, and challenges which lay primarily in the realm of bioinformatics (Gauthier *et al.*, 2019). Omics data can be generated on different levels in respect to the major biomolecule class in question, whether it be DNA (genomics), RNA (transcriptomics), protein (proteomics) or metabolite (metabolomics) (Figure 1.2).

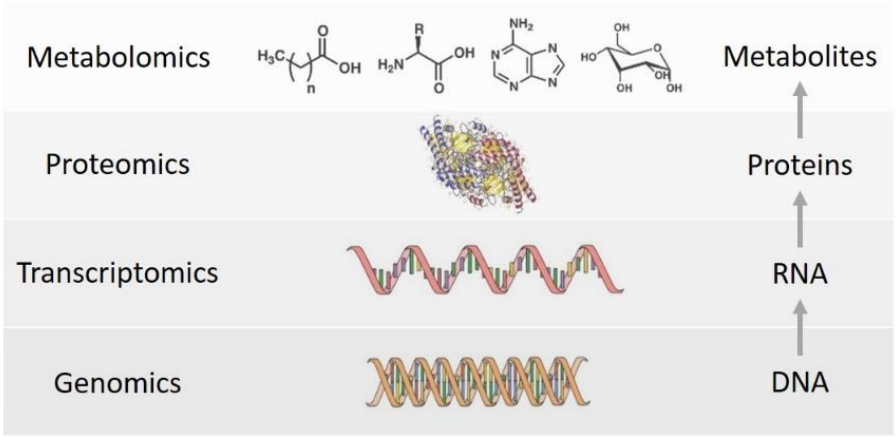


Figure 1.2: Different levels of omics based on the major biomolecules studied.

Genomics is defined as the study of complete sets of genetic material found within the cells of an organism, which not only includes both coding and noncoding DNA, but also the genetic material found in the mitochondria of most eukaryotic cells and chloroplasts in plant and algal cells. Genomics focuses on the study of whole genomes with respect to their structure and function, and the impact of variations within the genome on various aspects of life. Genomics data consist of the sequence of the DNA, which carries information ranging from single nucleotide variations to larger structural changes such as copy number variations, large deletions and insertions, and their subsequent annotation (Del Giacco and Cattaneo, 2012).

Transcriptomics is the term given to the qualitative and quantitative study of complete sets of transcripts, including coding and non-coding, within an organism (Chang, 2016). Due to the closer relationship between transcripts and proteins, and thus to the phenotype in comparison to the genome, transcriptomics is often the preferred level of omics to study cellular states such as differentiation and biomarker

discovery efforts. Phenomena such as alternative splicing and RNA editing are some of the difficulties that are faced upon studying whole transcriptomes. Current developments in the field of transcriptomics focus on studying the transcriptome of single cells. Such efforts aim to identify cellular subpopulations, determine whether detected changes are due to real cellular phenotypes or proliferation, investigate processes such as cellular differentiation and study rare populations of circulating tumor cells or cancer stem cells (Kanter and Kalisky, 2015; Trapnell, 2015; Chen, Ning and Shi, 2019).

Proteomics allows the study of entire proteomes and relative quantitative comparisons over various conditions with comprehensive proteome coverage (Mishra, 2010). It promises to provide a more complete description of the cellular state, since it informs the researcher about the end-point of the expression cascade, and the amounts and properties of proteins (Cox and Mann, 2011; Altelaar, Munoz and Heck, 2013; Aebersold and Mann, 2016). Two major types of proteomics strategies are bottom-up (Wolters, Washburn and Yates, 2001; Sinitcyn, Rudolph and Cox, 2018) and top-down proteomics (Marshall, 2006; Toby, Fornelli and Kelleher, 2016; Fornelli *et al.*, 2017). The bottom-up proteomics approach based on liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) techniques, often named shotgun proteomics, aims to measure digested proteins (peptides) since measuring intact proteins has proven to be cumbersome (Zhang *et al.*, 2013). Some of the difficulties in measuring intact proteins via mass spectrometry, so-called top-down proteomics, include highly complex spectra which are hard for deconvolution algorithms to handle due to the majority of ions within the sample being multiply charged, and the sample preparation hurdles that arise from having to deal with intact proteins, especially insoluble ones (Brown *et al.*, 2020). For bottom-up proteomics, the proteins are first digested using a protease. Trypsin is often the protease of choice because of its high fragmentation efficiency, suitable peptide length for HPLC separation and the fact that it cleaves peptides on the C-terminal side of lysine and arginine residues (both of which carry a positive charge), which is useful for the ionization of the peptides (Aebersold and Mann, 2003).

Metabolomics aims to study the entire set of small molecules, typically <1500 Daltons (Da) in mass, known as metabolites within an organism, tissue or cell (Weckwerth, 2007). Since metabolites are the omics level closest to the phenotype of an organism, the study of the metabolome is considered the closest one can get to the

function of the underlying biological mechanisms governed by genes, transcripts and proteins. Similar to proteomics, LC-MS/MS is the analytical platform of choice for metabolomics studies. LC-MS/MS-based technologies due to the increased sensitivity and potential for detecting novel unknown metabolites can cover a larger portion of the metabolome (Stringer *et al.*, 2016). The ultimate goal in metabolomics is to detect and quantify metabolites similar to expression analysis in transcriptomics or proteomics. By far the largest metabolomics data repository is the MetaboLights database (Haug *et al.*, 2013).

Genomics, transcriptomics, proteomics and metabolomics data is further divided into data that focuses only on a certain aspect of each of the aforementioned biomolecules. These areas of focus are typically the various chemical modifications that can be found on such biomolecules, which are known to have a functional significance. These include epigenetic modifications such as methylation (Rauluseviciute, Drabløs and Rye, 2019), alternative splicing in the case of the transcriptome (Ding, Rath and Bai, 2017), post-translational modifications (PTMs) on proteins (Larsen *et al.*, 2006) and structural variations in metabolites (Figure 1.3) (Dettmer, Aronov and Hammock, 2007). Several different analytical techniques and technologies are utilized for the generation of omics datasets, with the most popular being next generation sequencing (NGS) for genomics and transcriptomics, and LC-MS/MS for proteomics and metabolomics (Kandpal, Saviola and Felton, 2009). Proteomics and metabolomics based on LC-MS/MS still need further development to reach the level of genomics and transcriptomics, at both the technical and data analysis level (Smith *et al.*, 2014). In this section, the analytical platforms of choice for genomics and transcriptomics (NGS), and proteomics and metabolomics (LC-MS/MS) are discussed.

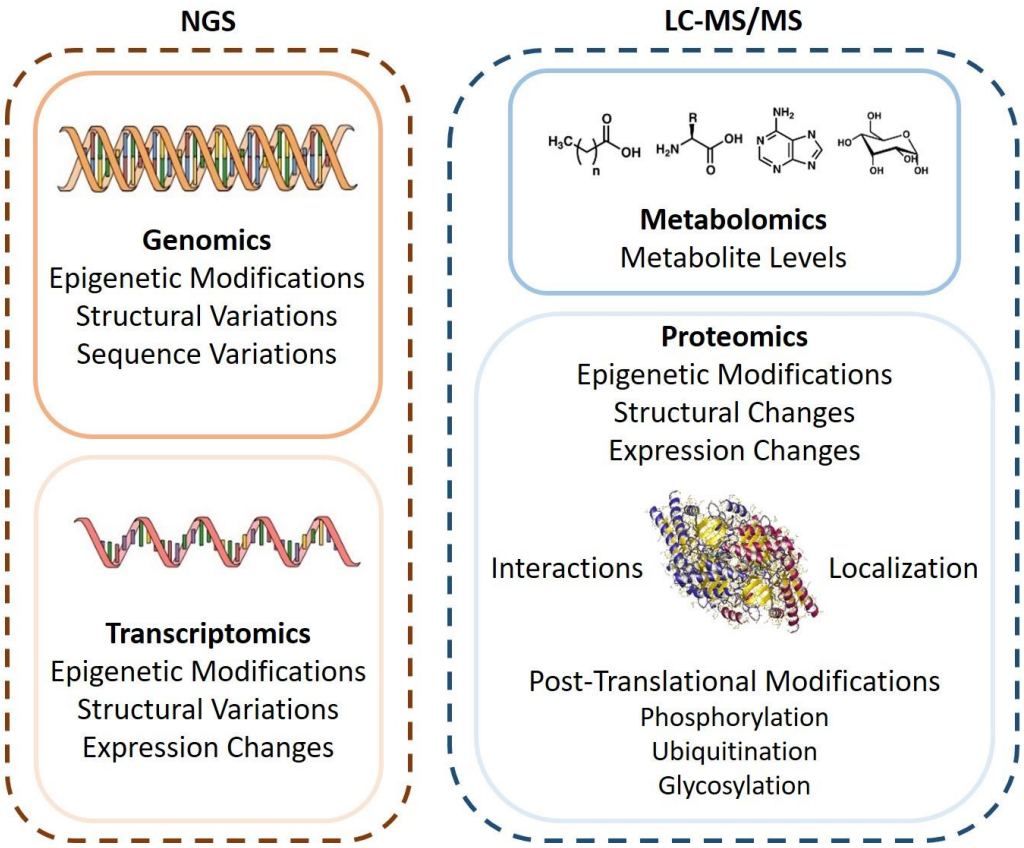


Figure 1.3: Various omics dimensions with two major analytical platforms, namely NGS and LC-MS/MS are shown. Each of the major omics data also have several subdivision such as secondary modifications in genomics, transcriptomics and proteomics, along with others such as localization in the case of proteomics.

1.1.1 Next Generation Sequencing

The analytical platform of choice for genomics and transcriptomics is NGS (Wang, Gerstein and Snyder, 2009; Koboldt *et al.*, 2013). It can arguably be considered the first spark in the data revolution in biological research. During the past two decades, scientists have been able to decipher the genome and transcriptome of many organisms from different domains of life using NGS technologies, which have now reached a level where improvements are incremental in both the hardware and software utilized (Giannopoulou *et al.*, 2019). It has quickly replaced microarrays and has been rapidly adapted to the clinic mainly due to ultra-high throughput, scalability, robustness and speed when compared to previous techniques such as Sanger sequencing or microarrays, paving the way for the increasing availability of personalized medicine (Hurd and Nelson, 2009). NGS empowers the average lab to sequence the entire human genome in less than 24 hours (Levy and Myers, 2016).

This is a huge feat when compared to Sanger sequencing which would require over a decade to deliver such data (Lander *et al.*, 2001) or microarrays, which are limited in their capacity to capture the entire genome on a single chip (Bumgarner, 2013). NGS owes its success largely to smart bioinformatics solutions, which are essential for constructing the genome in question from millions of fragments that are sequenced in parallel (Behjati and Tarpey, 2013). Whole genomes can now be easily sequenced and stored along with annotation information in publically available databases (Mailman *et al.*, 2007; Kodama, Shumway and Leinonen, 2012; Clough and Barrett, 2016). The sequencing of the human genome (Lander *et al.*, 2001) set the stage for omics studies of many kinds.

1.1.2 Liquid Chromatography Coupled with Tandem Mass Spectrometry

The most widely used analytical platform for proteomics and metabolomics is LC-MS/MS (Blum, Mousavi and Emili, 2018). It provides data in three dimensions, *m/z* (mass to charge ratio), retention time and intensity. Recent extensions of the LC-MS/MS setup (Figure 1.4), such as the FAIMS interface (ion mobility) (Hale *et al.*, 2020) add a fourth dimension to the data, which will not be covered here. In the classical LC-MS/MS setup, sample preparation, which may also include protein, peptide, metabolite or lipid fractionation and enrichment, is followed by high performance liquid chromatography (HPLC) and subsequently by mass spectrometry (MS) data acquisition. The generated raw data is then processed and analyzed to identify and quantify the features of interest (Sinitcyn, Rudolph and Cox, 2018).

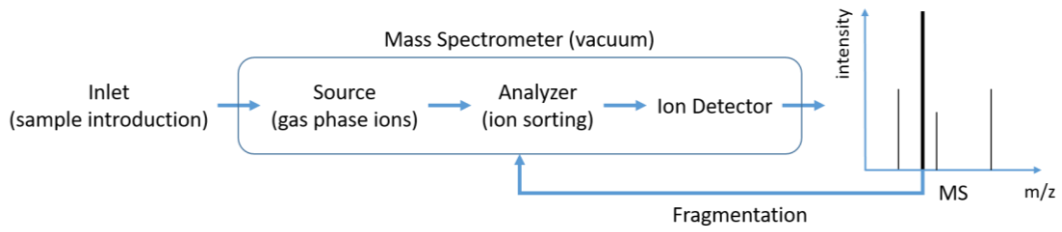


Figure 1.4: Basic LC-MS/MS setup.

Various fields of analytical chemistry utilize HPLC as an efficient method for separating, identifying and quantifying components in liquid mixtures (Dahimiwal *et al.*, 2013). The basic principle used in this technique is pumping liquid containing a desirable solvent and the sample in question through a column prepared with a

suitable solid adsorbent. The system then relies on varying interactions between the compounds in the liquid mixture and the adsorbent material, which would in turn alter the flow rate of the liquid mixture and thus separate the components going through the adsorbent packed column (Figure 1.5). The HPLC is used to derive the retention time of the ions, which is the time measured from sample injection to the HPLC and the appearance of the maximum signal for the ion post chromatographic separation (Katajamaa and Orešič, 2005). In the case of proteomics, purified proteins or peptides are separated after digestion with nanoliter per minute flow rates with the HPLC, prior to being introduced to MS analysis via electrospray ionization (Figure 1.6) (Hein *et al.*, 2013). It is estimated that complex digested proteome samples may contain well over a hundred thousand unique peptides (Michalski, Cox and Mann, 2011; Nagaraj *et al.*, 2011). Such complex samples cannot simply be resolved directly by MS and thus, HPLC is crucial for slower sample introduction, allowing the MS to be able to capture and measure as many peptides as possible. To this end, the sample is loaded to a chromatographic column, which is usually packed with a hydrophobic reverse phase material such as C18. This leads to the peptides binding the hydrophobic reverse phase material with different strengths based on their chemical properties and thus, be released gradually by increasing the amount of the solvent.

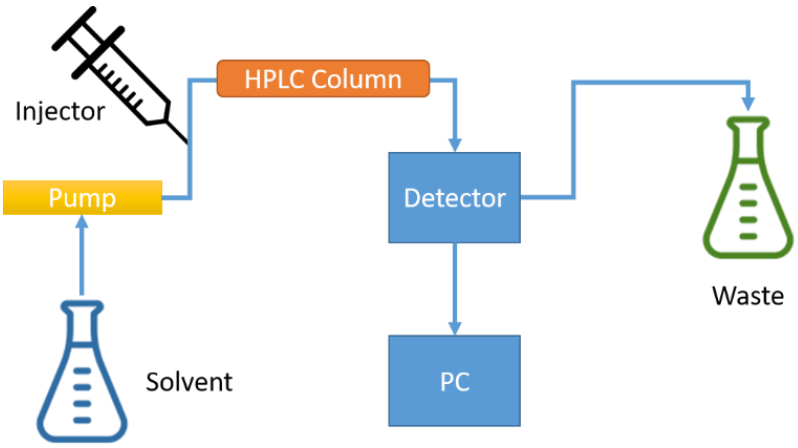


Figure 1.5: Basic schema of a HPLC setup.

The mass spectrometer has been a widely used platform in various field of research for measuring the *m/z* of ions. Measurements are often visualized as a mass spectrum where the intensity of an ion is plotted against its *m/z* ratio (Figure 1.6). Different techniques exist, which can be divided into two major groups, MS with trap-based mass analyzers and MS with beam-based analyzers. Regardless of the type of

mass analyzer, every mass spectrometer used in LC-MS/MS has three key elements, the ion source, the mass analyzer and the detector (Figure 1.4) (El-Aneed, Cohen and Banoub, 2009).

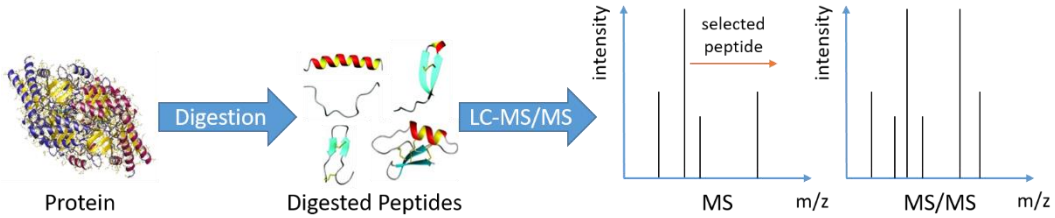


Figure 1.6: Basic bottom-up proteomics approach leading to MS and MS/MS mass spectra.

One goal in further developing new mass spectrometers is to reach a machine capable of higher resolution measurements. Resolution is the ability to distinguish two features corresponding to two different ions with very similar *m/z* ratio (Scigelova *et al.*, 2011) and is calculated as the ratio of a features' *m/z*, and the delta *m/z* at the full width half maximum (FWHM) of that feature (Marshall and Hendrickson, 2008). Another goal is to reach better mass accuracy, which is the deviation between the theoretical mass and the experimentally determined mass of ion. Mass accuracy is influenced by the resolution of the mass spectrometer and the signal to noise ratio within the MS data (G. Marshall *et al.*, 2013). Furthermore, it is important to have a mass spectrometer that not only detects highly abundant ions, but also capable of detecting low abundant ions in complex mixtures. This is known as the dynamic range. The scan speed is also important as it defines how fast a certain *m/z* range can be scanned, which is mostly inversely correlated with the resolution of the mass spectrometer (Wu and Han, 2006). Finally, new mass spectrometers aim for higher sensitivity, which is measured by the intensity of the MS signal for a certain concentration of the sample. Many different operation modes exist for mass spectrometers. In its targeted mode, a predefined target mass range is set with the aim of reaching the highest possible quantitative accuracy and reproducibility (Marx, 2013). On the other hand, DDA and DIA mass spectrometry (Zhang *et al.*, 2013, 2020) aim to capture the largest possible spectrum, both of which will be discussed later in the chapter.

The latest game-changing mass spectrometer technology introduced is the Orbitrap mass analyzer (Hu *et al.*, 2005; Olsen *et al.*, 2005; Michalski *et al.*, 2011),

which is similar to Fourier transform ion cyclotron resonance mass spectrometers (FTMS) in its working principle, where ions are trapped using an electrostatic force and thus orbit around a small spindle shaped electrode (Figure 1.7). This electrode is designed in a manner that the orbiting ions are not only confined in their orbit, but also oscillate along the length of the electrode. The oscillation is used to obtain what is called an image current in the detector plates of the Orbitrap, which is subsequently recorded by the mass spectrometer. Since the frequencies of the image currents are related to the *m/z* ratios of the ions, the relevant mass spectra can be obtained from performing Fourier transformation on them (Hu *et al.*, 2005; Scigelova *et al.*, 2011). The Orbitrap is with its innovative working principle is the latest addition to the array of different types of mass analyzers that are utilized in modern mass spectrometers. It was introduced almost two decades ago (Makarov, 2000), and has been quickly adopted by biologists as the go to platform for proteomics and metabolomics studies. Its success is due to higher resolution, high mass accuracy and a large dynamic range (Hu *et al.*, 2005; Makarov *et al.*, 2006).

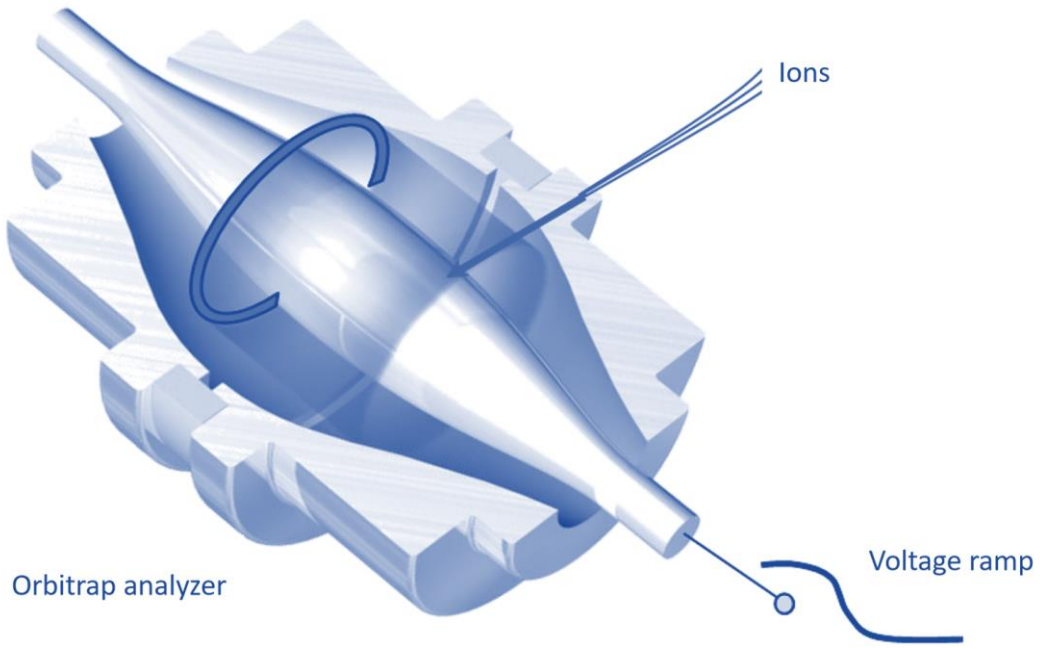


Figure 1.7: The structure of the Orbitrap mass analyzer.

1.2 Computational Proteomics

Advancements in proteomics technologies have been rapid and thus, effective in earning proteomics a significant place in today's biomedical research (Cox and Mann,

2008, 2011; Aebersold and Mann, 2016), but many areas are still in need of further development. These include computational methods for data processing and analysis (C. Chen *et al.*, 2020) and some of the most important advancements in this direction are computational platforms and workflows such as MaxQuant and Perseus (Cox and Mann, 2008; Cox *et al.*, 2011; Tyanova *et al.*, 2016; Sinitcyn *et al.*, 2018). MaxQuant is a software suit, which provides an easy and intuitive means for performing quantitative proteomics data analysis for large LC-MS/MS data sets, and Perseus provides an intuitive and user-friendly platform for the downstream analysis of MaxQuant outputs.

Peak detection within the spectra generated by MS is an important initial step in computational proteomics (Zhang *et al.*, 2009), and with the introduction of higher resolution MS machines, it has been possible to resolve the isotope pattern and even fine structures of peptides (Miladinović *et al.*, 2012). The peak information (*m/z* and intensity) coupled to the retention time information from HPLC become 3D features (Figure 1.8). These features are then taken and assembled to construct isotope patterns. This information when combined, lead to considerably high mass precision, but this is not necessarily true for mass accuracy, primarily due to systematic errors that occur during MS measurements. Such errors have been observed to be typically nonlinear and dependent on *m/z*, retention time, and signal intensity. In case of LC-MS/MS coupled with ion mobility, the ion mobility index also has an effect on the mass error (Sinitcyn, Rudolph and Cox, 2018). MaxQuant was first introduced with an effective algorithm for tackling the mass error problem using a multivariate nonlinear recalibration algorithm, which takes advantage of the many peptides within complex proteomics samples as calibration points, resulting in significant increases in mass accuracy (Cox and Mann, 2008, 2009; Cox, Michalski and Mann, 2011). Besides mass accuracy, in order to ensure consistent retention time values and ion mobility orders for peptides across different runs, similar recalibration strategies are employed. This is important since HPLC and ion mobility setups are naturally prone to irreproducibility, causing problems when comparing different LC-MS/MS runs (Sinitcyn, Rudolph and Cox, 2018). Following these recalibration steps, it is possible to transfer identifications across different runs and compare several different runs together (Paša-Tolić *et al.*, 2004), which is especially useful to tackle the stochastic nature of DDA methods in peptide fragmentation (Tyanova, Temu and Cox, 2016).

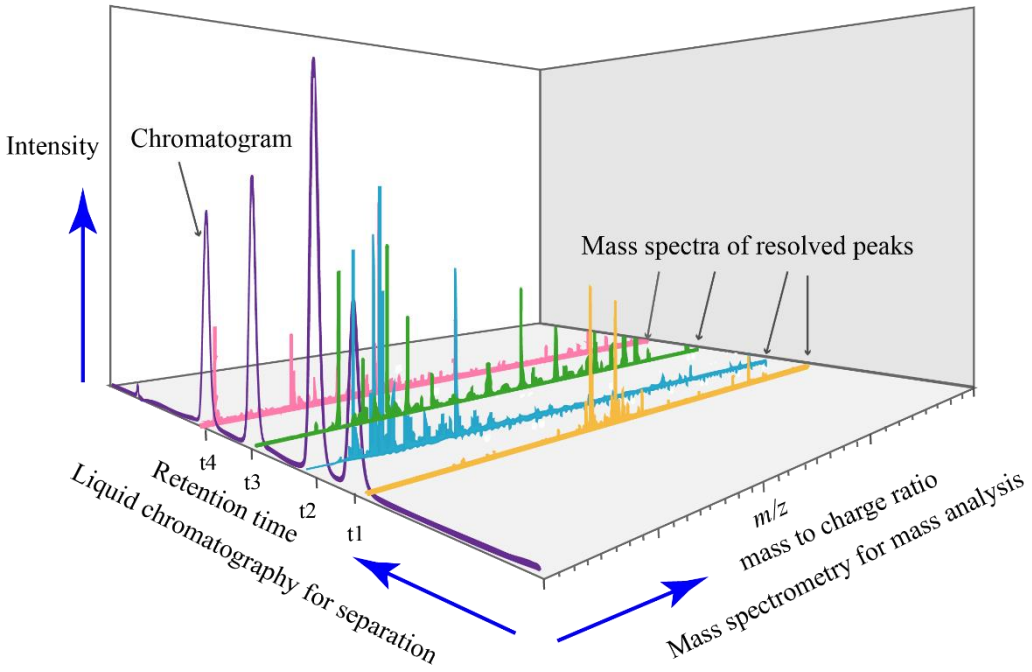


Figure 1.8: 3D peak information.

The fragmentation spectra obtained after the initial MS scan are analyzed in order to sequence the peptides. This is done most frequently using a database search engine approach, where the database contains all theoretical peptide fragmentation spectra generated *in silico*, using whole genome information (Craig and Beavis, 2004; Geer *et al.*, 2004; Cox *et al.*, 2011). Such approaches match the measured fragmentation spectrum to the entries within the database considering a certain mass tolerance. Each such match is named a peptide-spectrum match (PSM), and to control for false positive PSMs, a target-decoy approach is employed (Elias and Gygi, 2007), where in addition to the target database of all theoretical peptide fragmentation spectra, a decoy database is constructed. The decoy database often contains the reverse sequences of the target database and matches are labeled as false-positive PSMs. The score distributions of the PSMs from the target and decoy databases can then be used to calculate posterior error probabilities, and control for the false discovery rate (FDR) along with other peptide features such as peptide length and number of missed cleavages (Cox and Mann, 2008). After peptide identification, the peptides are assembled into proteins. The main challenge in this step is that during protein digestion many peptides are digested from a protein and many peptides are not unique to a certain protein, thus, there is a many-to-many relationship between peptides and proteins (Huang *et al.*, 2012).

After peptide identification and protein assembly, the proteins can then be quantified. Protein quantification can be either absolute or relative. Absolute quantification aims to determine the quantity of a protein within a certain sample, whereas relative quantification deals with determining the ratio of the protein quantity between samples (Figure 1.9). Quantification strategies can be based on using labels, e.g. using stable isotopes to tag peptides, or be performed in a label-free manner. Relative label-free quantification (LFQ) is challenging due to the nature of LC-MS/MS data. These challenges include retention time differences between LC-MS/MS runs due to parallel sample handling and irreproducibility of HPLC, stochastic MS/MS sequencing as the mass spectrometer chooses the most abundant peptides for MS/MS leading to missing peptide identifications across samples, and pre-fractionation causing peptides to appear in several fractions. MaxQuant, equipped with the MaxLFQ algorithm, overcomes such challenges via nonlinear retention time alignment, peptide identification transfer between different runs and peptide intensity normalization across fractions (Cox *et al.*, 2014). In term of absolute quantification, MS is not inherently quantitative due to the vastly different behavior of peptides within the mass spectrometer, and the strong correlation of the MS signal with the input amount of the protein. To overcome these challenges, Perseus is equipped with the Proteomic Ruler plugin, which uses the histone signals identified within the MS run as a scale with respect to the amount of DNA measured in the sample, to estimate protein copy numbers (Wiśniewski *et al.*, 2014).

	S.1	S.2	S.3	S.4	S.5
Prot.1	3	4	1	3	5
Prot.2	1	0	9	5	6
Prot.3	2	3	0	7	7
Prot.4	5	7	4	8	8
Prot.5	3	4	3	5	1

Figure 1.9: Absolute and relative protein quantification. The vertical yellow shaded bar depicts absolute quantification within a sample and the horizontal blue shaded line depicts relative quantification across different samples.

Computational proteomics has been a corner stone of proteomics studies (Sinitcyn, Rudolph and Cox, 2018). Its ultimate goal is to process and analyze the data generated primarily via LC-MS/MS to identify and quantify proteins for

studying comparative changes between different conditions, posttranslational modifications, protein-protein interactions, and the subcellular localization of proteins. Generally, endeavors in computational proteomics can be divided into two major groups, the correct identification and precise quantification of proteins, and the detailed analysis of this information within the context of the specific biological question. To this end, bioinformaticians have to develop various algorithms for handling the different types of LC-MS/MS data acquisition methods. In the following two sections, DDA and DIA proteomics approaches will be discussed.

1.2.1 Data Dependent Acquisition

One of the most mature acquisition methods in proteomics is data dependent acquisition (DDA) (Dupree *et al.*, 2020). It is the most widely adapted method for proteomics studies, in which ions are separated after the first MS scan based on their *m/z*, and the instrument then selects certain ions in real-time with specific *m/z* values for further analysis after fragmentation (MS2) (Figure 1.10). Subsequent fragmentation techniques post precursor ion selection vary, and can be any of collision-induced dissociation, ion-molecule reaction, or photo-dissociation.

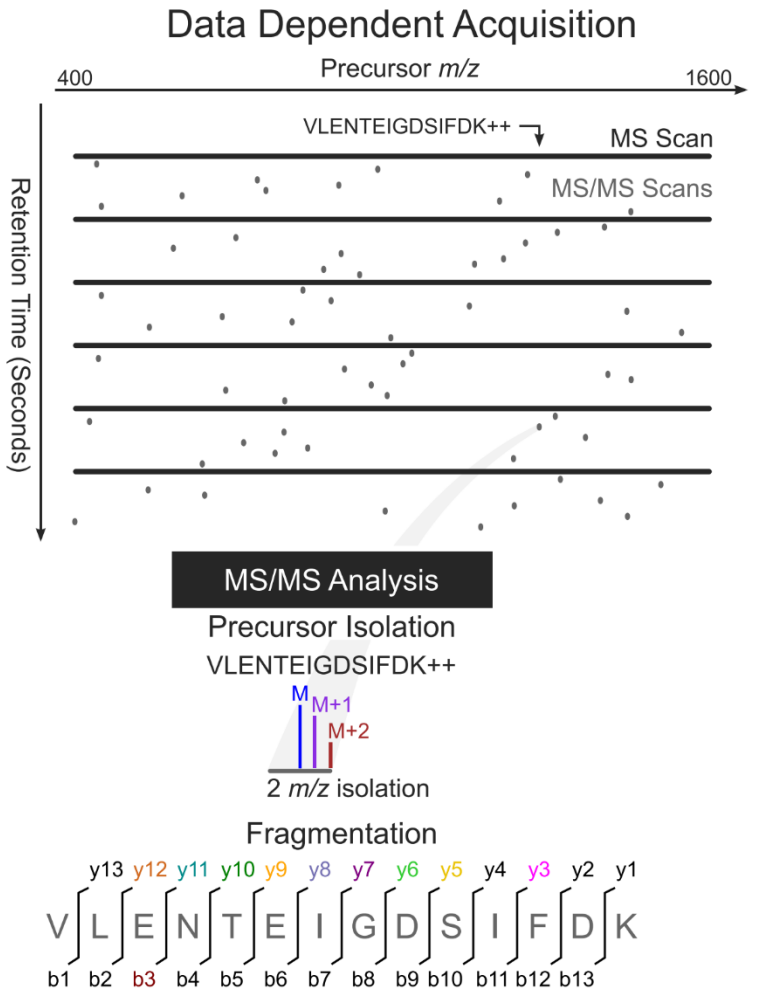


Figure 1.10: Schematic overview of data dependent acquisition proteomics (adapted from (Wolf-Yadlin, Hu and Noble, 2016)*).*

Using DDA, the mass spectrometer is set to select a subset of ions based on MS1-level data, usually the most abundant ions, for further analysis via MS2. This is why DDA is also sometimes named the topN method (Venable *et al.*, 2004). Higher abundant ions are preferred, since they usually lead to higher quality MS/MS spectra, leading to a higher number of identifications (Hebert *et al.*, 2018). In this step, the mass spectrometer uses ion fragmentation and tandem measurement to deliver further information on the ion in question. In the case of peptides, the fragmentation energies are set in a way that they are most optimized for single peptide backbone breakages, leading to a set of complementary fragment ions. DDA has improved with each new generation of mass spectrometers to capture more ions, leading to efficient capture of effectively a complete proteome even with a single run MS run (Bekker-Jensen *et al.*, 2017).

1.2.2 Data Independent Acquisition

The LC-MS/MS method for studying proteins, especially in its DDA form for shotgun proteomics has allowed for the in-depth analysis of entire proteomes. In efforts to make proteomics techniques a more stable and robust, new methods are being developed using DIA. DIA promises to be a faster, thus cheaper alternative to the current gold standard that is DDA. In DIA, all ions within a selected *m/z* range are sent for fragmentation for further analysis in the second MS (Figure 1.11). It has been in use and constant development during the last two decades, and has continued to be utilized and improved, with methods focusing on fragmenting entire precursor ranges, also with narrower windows which aim to simulate DDA runs (Panchaud *et al.*, 2009, 2011; Geiger, Cox and Mann, 2010b; Egertson *et al.*, 2013). Fragmenting entire precursor ranges result in faster data acquisition and thus, help to cover wider mass ranges at the cost of higher spectral complexity, and utilizing narrower windows lead to less complex spectra with a higher dynamic range at the cost of higher cycle times (Chapman, Goodlett and Masselon, 2014). DIA is progressively attracting traction for proteomics studies as it promises the advantages of targeted approaches to studying complete proteomes, especially in terms of sensitivity and reproducibility (Doerr, 2014). DIA strives to overcome the limitation in the number of MS/MS spectra that the mass spectrometer is able to measure by isolating certain ions. Since in DIA instead of a certain ion, a *m/z* range is selected for further fragmentation and analysis, the resulting MS/MS spectra is essentially a combined spectrum for multiple peptide precursors, which would need to be deconvoluted for effective peptide identification (Masselon *et al.*, 2000). On the other hand, DIA ensures that essentially no data is lost and all precursors are fragmented and thus, it not only promises to capture and record the entire proteome in the realm of the mass spectrometers maximum dynamic range, but also allows for higher reproducibility across different samples (Gillet *et al.*, 2012). Although in theory DIA has been proposed to have many advantages over the current DDA approaches, in practice it has so far not been able to compete with DDA in terms of whole proteome coverage (Röst *et al.*, 2014; Navarro *et al.*, 2016; Collins *et al.*, 2017). Perhaps this is mainly due to the lack computational workflows, which can effectively decipher the data to reach higher rates of identification and reliable quantification. Initial computational solutions for the analysis of the data were focusing on generating so called pseudo-MS/MS spectra, where fragment ions were grouped based on retention time information, and makeshift use of search engines initially designed for DDA data

(Bilbao *et al.*, 2015). MaxQuant equipped with MaxDIA, utilizes two different strategies for the analysis of DIA data based on both experimentally generated libraries using DDA methods and predicted libraries, which is further discussed in detail in section 4.2.

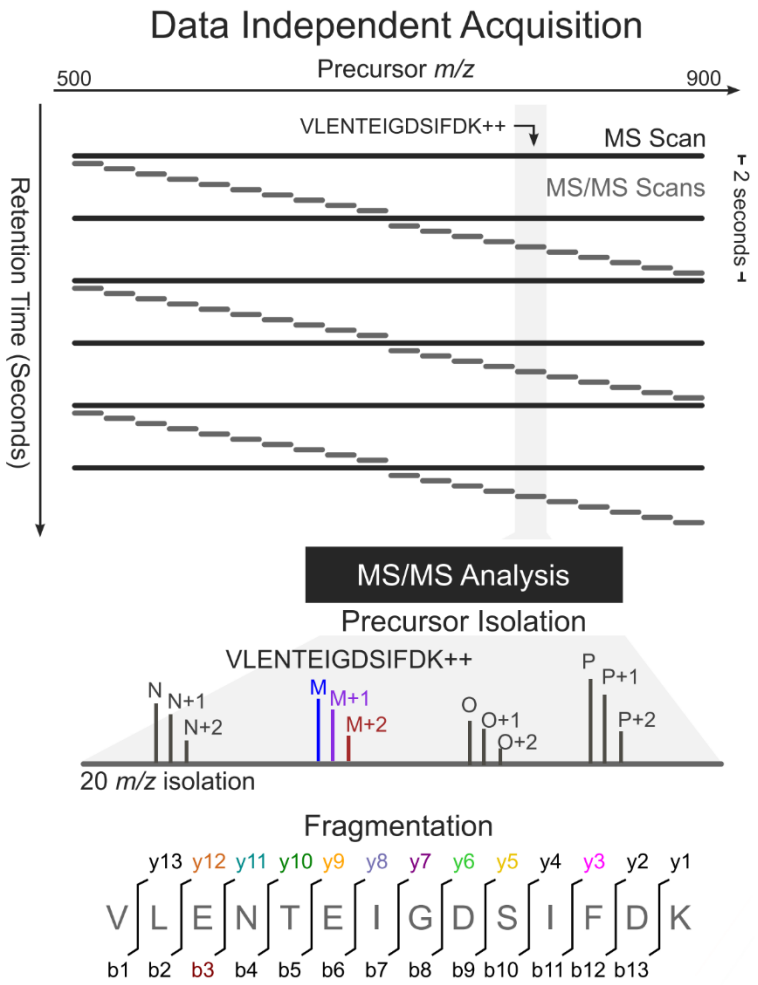


Figure 1.11: Schematic overview of data independent acquisition proteomics (adapted from (Wolf-Yadlin, Hu and Noble, 2016)*).*

1.3 Computational Metabolomics

The latest LC-MS/MS platforms can measure well over 200,000 ions within each run from typical biological samples, from which only a fraction of a percent are identified. Computational workflows for such metabolomics data aim to reach higher mass accuracy and effectively use information such as chromatographic retention time, collision-induced dissociation products and collision cross section for metabolite identification (Uppal *et al.*, 2016). Numerous tools exist for the processing of

metabolomics data designed to deal with the various aspects of the complexity of the data ranging from peak detection to spectral noise removal and feature alignment between different runs (Tautenhahn, Bottcher and Neumann, 2008; Yu *et al.*, 2009; Pluskal *et al.*, 2010).

Peak detection is done in a per file fashion with certain criteria such as signal-to-noise ratio and peak shape used to filter for peaks of high quality. This is followed by alignment strategies to create a dataset of all peaks contained within all files so to compensate for deviations and errors across LC-MS/MS runs. Such deviations and errors occur especially within the retention time dimension, which could arise from HPLC variables such as column temperature and the pressure within the system along with the changes within the column during the course of the runs (Lange *et al.*, 2008). Deviations could also arise during mass spectrometry and technical replicates are important to detect and account for such deviations (Uppal *et al.*, 2013; Libiseller *et al.*, 2015). One of the most important factors in metabolomics data analysis is mass accuracy. Since mass accuracy has a direct influence on the quality of alignment between samples, downstream feature annotation and metabolite identification, low mass accuracy jeopardizes the entire analysis (Kind and Fiehn, 2006). To this end, mass error correction strategies exist that exploit internal standards and references, which can estimate the error, and account for it downstream for improving alignments between different runs (Shahaf *et al.*, 2013). Such strategies are effective to a degree, but due to the limited amount of standards and references, it is often difficult to account for the mass error across the entire mass range of the measurements. The following section discusses these limitations along of with some lessons from proteomics and possible solutions.

1.3.1 What Computational Non-Targeted Mass Spectrometry-Based Metabolomics can gain from Shotgun Proteomics

In the following review article (Hamzeiy and Cox, 2017), the common challenges between computational proteomics and metabolomics are discussed with a focus on how these challenges are met in the realm of proteomics, and how such strategies can be adapted to the field of metabolomics. It is argued that similar to the effect of higher mass accuracies in proteomics datasets where higher rates of identification are achieved, metabolomics datasets would also benefit from a smart mass recalibration algorithm, with the end goal of reaching higher rates of metabolite identification.

Contributions to the following review within the context of this thesis include the gathering and organizing of all publicly available metabolomics data for preliminary testing of the mass recalibration strategy proposed for metabolomics datasets, taking part in the implementation of the algorithm, testing and benchmarking, and writing the review.

Hamzeiy, Hamid, and Jürgen Cox. 2017. "What Computational Non-Targeted Mass Spectrometry-Based Metabolomics Can Gain from Shotgun Proteomics." Current Opinion in Biotechnology 43: 141–46. https://doi.org/10.1016/j.copbio.2016.11.014.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

What computational non-targeted mass spectrometry-based metabolomics can gain from shotgun proteomics

Hamid Hamzeiy and Jürgen Cox

Computational workflows for mass spectrometry-based shotgun proteomics and untargeted metabolomics share many steps. Despite the similarities, untargeted metabolomics is lagging behind in terms of reliable fully automated quantitative data analysis. We argue that metabolomics will strongly benefit from the adaptation of successful automated proteomics workflows to metabolomics. MaxQuant is a popular platform for proteomics data analysis and is widely considered to be superior in achieving high precursor mass accuracies through advanced nonlinear recalibration, usually leading to five to ten-fold better accuracy in complex LC–MS/MS runs. This translates to a sharp decrease in the number of peptide candidates per measured feature, thereby strongly improving the coverage of identified peptides. We argue that similar strategies can be applied to untargeted metabolomics, leading to equivalent improvements in metabolite identification.

Address
Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Martinsried, Germany

Corresponding author: Cox, Jürgen (cox@biochem.mpg.de)

Current Opinion in Biotechnology 2017, **43**:141–146

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Jurre J Kamphorst** and **Ian A Lewis**

For a complete overview see the Issue and the Editorial

Available online 28th December 2016

http://dx.doi.org/10.1016/j.copbio.2016.11.014

Introduction

Mass spectrometry-based proteomics [1,2,3••] has matured during recent years to a degree that makes it readily usable as a standard research tool in many branches of biological and biomedical research. Most often proteomics is implemented in the form of shotgun proteomics, in which proteins are first digested to peptides, separated by liquid chromatography, and finally studied in the mass spectrometer as intact peptides as well as their fragmentation patterns (LC–MS/MS). Proteomics applications include expression proteomics, analysis of protein–protein interactions [4], study of post-translational modifications [5], as well as determination of subcellular localization [6], which can all be done in a dynamic, time-dependent manner [7]. Most proteomics experiments can be performed without the use of labels, owing to appropriate algorithms for relative label-free quantification [8]. The complete yeast proteome can be quantified nowadays with moderate effort and studied in many different conditions [9–11], while in human cellular proteomes a depth of 10,000 proteins can be achieved [12–15].

Ten years ago, the situation was very different. Proteomics projects were very time consuming since the data analysis was mostly done in a semi-manual fashion. While peptide database search engines [16–20] and other software and algorithms for the identification and quantification of peptides already existed in principle, a lot of manual validation was still necessary in order to obtain reliable results that could be used for solid biological interpretation.

Certainly, technological improvements like the introduction of the Orbitrap mass spectrometer [21–23] and improvements in sample preparation also contributed to today's proteomics workflows to be evermore robust and easy to use. However, a large part of the improved situation is owed to the software platforms and computational workflows that have become mature and reliable. This starts with basic activities such as feature detection, correct label assignment, and processing of MS/MS spectra. Then, the identification process can reliably be controlled by false discovery rates on the peptide-spectrum match (PSM) or protein level. Furthermore, the results of quantification methods became better than what could be achieved with manual analysis. All these improvements together lead to a situation in which shotgun proteomics data analysis is approaching a state of maturity that is comparable to next generation sequencing data analysis. Also, software tools that aid in the biological interpretation of quantitative proteomics results are available and well accepted in the community [24].

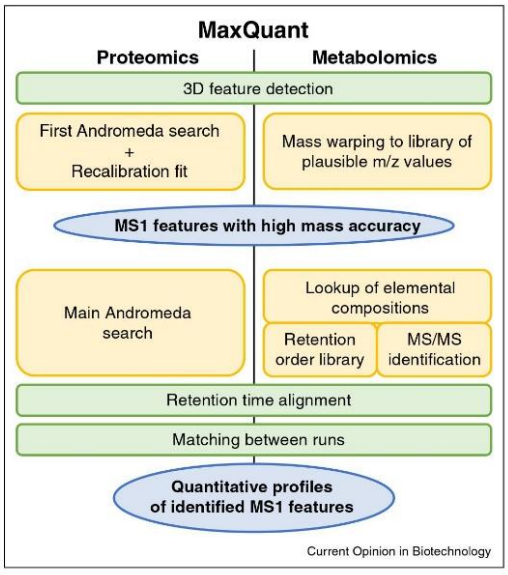
One of these computational workflows is MaxQuant [25,26•], including the Andromeda peptide search engine [27], which provides a complete solution for most standard quantitative experimental designs in shotgun proteomics. Its development provided seminal contributions to the reliable automation of the data analysis workflow. One aspect in which MaxQuant is unique is how it improves the mass accuracy of peptide features using computational techniques [28,29]. Nonlinear mass recalibration is applied to the MS1 features in an m/z and retention time dependent way. Multiple mass measurements over

elution profiles and isotopic peaks are then integrated, achieving mass accuracies in the ppb range for standard Orbitrap data in a complex proteomics run, which is a 5–10-fold increase over standard techniques.

Untargeted metabolomics [30,31] is a highly evolved field with many applications already accessible and high promises for the future. A wealth of analytical techniques [32] exist for its study and many computational tools [33,34] have been developed within the community. However, interpreting mass spectrometry-based untargeted metabolomics data remains a challenge and limits the translation of results into biologically relevant conclusions [35••]. Although the power of untargeted profiling is undeniable, it is the case that most mechanistic links are still revealed by hypothesis-driven targeted methods [36•]. This is likely due to untargeted metabolomics typically yielding complex data patterns that are not easily amenable to intuitive interpretation [36•]. One could make the provocative statement that untargeted metabolomics is several years behind shotgun proteomics in terms of ease of data analysis and interpretation.

Our plan is to create a version of MaxQuant for the analysis of untargeted metabolomics LC–MS data whose workflow follows loosely the shotgun proteomics workflow as sketched in Figure 1. Several important data processing steps can be transferred to metabolomics with only minor adaptations, as for instance the 3D feature detection, retention time alignment and matching of features between LC–MS runs based on accurate masses and retention times. Other processing steps need more changes in order to become applicable to metabolomics data. For instance, the two-level peptide search strategy, in which the identifications of the first round of searches are used to determine multidimensional nonlinear recalibration curves need to be replaced by another two-level strategy based on mass warping due to the absence of a universal search engine approach in metabolomics. We strongly believe that the application of this sort of nonlinear mass recalibration to metabolomics data is highly beneficial for compound identification by increasing the range of molecules for which an elemental composition can be assigned.

Figure 1



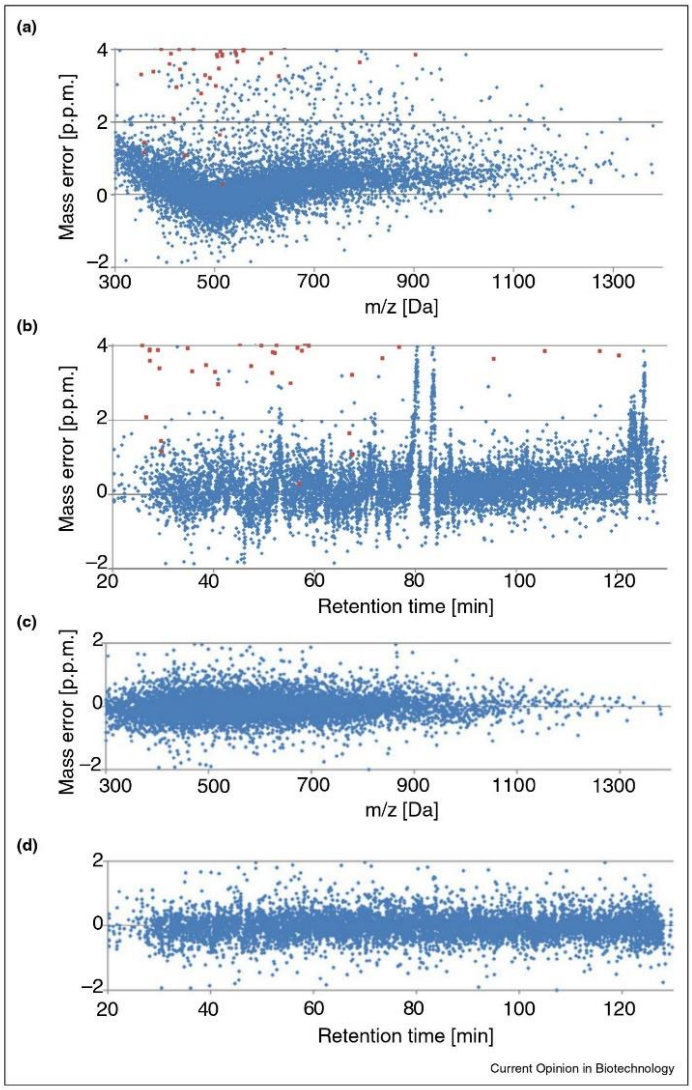
Schematic overview of high mass accuracy feature identification and quantification workflows in MaxQuant for shotgun proteomics and for untargeted metabolomics.

Improvement of mass accuracy in proteomics

Here we describe how high mass accuracy is achieved by mass recalibration algorithms in proteomics. In the next section we sketch our path to implementing similar improvements in untargeted metabolomics. For the determination of the nonlinear mass recalibration curves in proteomics we follow a strategy employing two consecutive peptide database searches (Figure 1). After having performed the 3D feature detection, a first round of Andromeda searches is performed. The purpose of this search is to generate a list of features with known masses which can then be used for recalibration. The precursor mass tolerance for the first search is relatively large, for example, 20 ppm, to be able to also correct for larger instrumental drift. Since there are many peptides available in a complex shotgun proteomics run, we can be restrictive at this stage and accept only identifications that are correct with high certainty, for example, by requiring a high Andromeda score threshold, which will typically still result in thousands of peptides per LC–MS run. Alternatively, one can use standards instead of the first search identifications. However, this strategy has the disadvantage that only a few features of known mass are available, which is usually not sufficient to perform the nonlinear recalibration to the accuracy attainable through the approach using many peptides from the sample. Figure 2(a) shows the mass deviations in a typical LC–MS run as a function of m/z, while in Figure 2(b) they are shown as a function of retention time. Clearly, just linear recalibration would leave many mass deviations far above 1 ppm.

Once we have obtained the long list of known masses, we fit a model to the mass deviations describing them by nonlinear dependencies on m/z and retention time. For time-of-flight mass spectrometers, the intensity dependence of the mass error is estimated and corrected (not necessary for Orbitrap data). No particular functional form of these dependencies is assumed. Instead, we use either splines or piecewise linear functions as models for the m/z and retention time dependencies. Overfitting is avoided

Figure 2

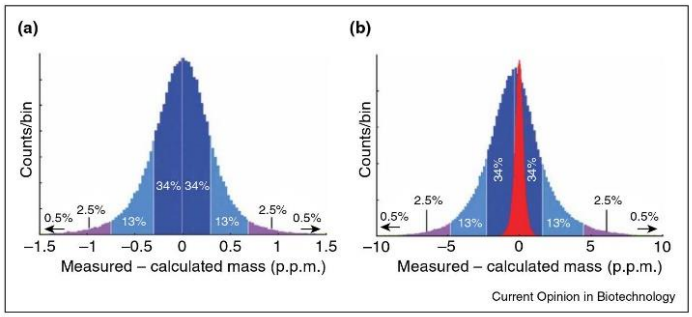


Nonlinear mass recalibration in MaxQuant. **(a)** *m/z* dependence of the mass error before recalibration on an Orbitrap mass spectrometer. **(b)** Retention time dependence of the mass error before recalibration on an Orbitrap mass spectrometer. **(c)** and **(d)** Same as **(a)** and **(b)** after application of the nonlinear mass recalibration functions.
Adapted from Ref. [28].

by keeping the ratio of number of parameters to number of data points to a low percentage number. Figure 2(c–d) shows the residuals after applying the calibration functions to the data which fluctuate independently around zero.

After having obtained the nonlinear recalibration functions, these are applied to all peptide features, also to the ones that were not used in the fit. This includes those MS1 feature that were fragmented, but not included in the

Figure 3



Mass error distributions before and after nonlinear mass recalibration. The red histogram on the right side is the same as the histogram on the left side and was added for comparison.
Adapted from Ref. [25].

recalibration fit due to the Andromeda score threshold. It also includes the MS1 features that were not fragmented at all, which is usually a vast majority of the signals [37]. Figure 3(a) shows a histogram of mass deviations obtained after recalibration in a typical LC–MS run. The average absolute mass deviation (absolute value of the difference between measured and calculated masses) is below 300 ppb. In Figure 3(b) the same histogram is recorded for masses taken directly from the instrument software without applying MaxQuant recalibration. Since the histogram is centered near zero, a linear shift as recalibration would obviously not improve the mass accuracy much. For comparison purposes, the red histogram was included which is the same as in Figure 3(a). For this typical LC–MS run the mass accuracy was improved by about 6-fold using MaxQuant recalibration routines.

Such a strong increase in mass accuracy will have implications on the peptide identification process. When searching in the human proteome the corresponding shrinkage of the precursor mass tolerance window — which is individualized for each peptide in MaxQuant — will translate proportionally into a restriction of possible peptide candidates for a given MS1 feature. Therefore, less information needs to come from the MS/MS spectrum to have the same certainty of identification of a peptide. With a fixed false discovery rate of, for example, 1% for PSMs (and 1% for proteins) the coverage of identified proteins will rise [29]. The extent of the improvement depends on many factors, like size of the protein sequence space used for generating the in silico peptide list for the database search, the type of digestion and the number of variable modifications.

Improvement of mass accuracy in metabolomics

Similar concepts as described in the previous section can be applied to non-targeted metabolomics. While our work in proteomics is mostly agnostic of the mass spectrometric instrument, here we focus on the Orbitrap since the scaling of resolution with the mass range is favorable for small masses. A central part of the proteomics workflow is the generation of MS1 features with known masses through the 'first Andromeda search' (Figure 1). In principle, one could follow a similar route and replace the peptide database search engine with a spectral library search and accept only indisputable identifications. However, we decided to adapt a different strategy that would also be applicable in the absence of MS/MS data.

We first generate a library of 'plausible m/z values' that one is likely to find in a metabolomics LC–MS run. This is in the first instance filled with all molecules from databases of compounds with biological relevance, such as CheBI [38]. Then we perform the MaxQuant 3D feature extraction on a large amount of untargeted metabolomics LC–MS runs in order to find which of the features can be interpreted as an adduct of a molecule that is already in the library of plausible m/z values, which are then also added to the library of 'plausible m/z values. The library contains all isotopic peaks, not only monoisotopic masses, since the subsequent algorithms will work on the 3D peak features before assembling them to isotope patterns.

Each LC–MS run to be analyzed is then mass aligned to this list of plausible m/z values. For this we use a kind of warping algorithm that finds an optimal nonlinear calibration function under the objectives as bringing as many MS1 features as possible as close as possible to a value in the list of plausible masses. This is done while requiring smoothness of the recalibration function in order to avoid overfitting. In this optimization procedure most of the MS1 features will 'snap' to the correct elemental composition. Some will not, because the correct composition is

not present yet in the library. The algorithm will still be able to find a good interpolating solution due to the smoothness requirement.

The library is a dynamic entity which will be updated based on the knowledge gain resulting from each alignment with an LC–MS run. If, after a new alignment there are unmatched MS1 features left with good signal-to-noise, and fitting a plausible new elemental composition, it will be added to the library. The alternative to this procedure would be to work in the space of all theoretically possible elemental compositions. However, we think there are big advantages to build up this reference list bottom up from real data and not have it filled up from the beginning with things that will never be seen in actual LC–MS runs.

The degree to which mass accuracy helps in reducing the number of possible molecular formulas depends on many factors, including the molecular mass and assumptions on the space of possible formulas. Under reasonable assumptions the number of candidate formulas shrinks considerably when going from 5 ppm to sub ppm accuracy over a wide range of masses as shown in Table 3 of Ref. [39]. Orthogonal filters like isotopic abundance ratios or ion mobility measurements would certainly diminish the number of candidates as well. Preliminary results show that the increase in mass accuracy obtained by our proposed method is indeed comparable to the gains seen in shotgun proteomics. The resulting reduction in candidates will lead to complete determination of elemental compositions for the majority of MS1 features. This will improve MS1-only workflows that use a lab-specific retention order library for distinguishing isomers. Metabolic flux [40–43] analysis can be supported as well by including the ^{13}C-labeling patterns of metabolic intermediates or end products into the list of plausible m/z values.

Conclusions

The adaptation of MaxQuant to untargeted metabolomics will strongly improve the mass accuracy of MS1 features. Similar to proteomics, this increased identification information will strengthen the robustness of the automated data analysis workflow in untargeted metabolomics. Together with other features from the MaxQuant workflow that are readily transferable to metabolomics — retention time alignment and matching between runs — MaxQuant should yield a useful addition to the computational metabolomics toolbox.

Conflict of interest statement

The authors declare no competing financial interests.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 686547.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Cox J, Mann M: **Quantitative, high-resolution proteomics for data-driven systems biology**. *Annu Rev Biochem* 2011, **80**:273-299.

2. Altelaar AF, Munoz J, Heck AJ: **Next-generation proteomics: towards an integrative view of proteome dynamics**. *Nat Rev Genet* 2013, **14**:35-48.

3. •• Aebersold R, Mann M: **Mass-spectrometric exploration of proteome structure and function**. *Nature* 2016, **537**:347-355.
This is a recent review of mass spectrometry-based proteomics summarizing its achievements and the remaining challenges. It is summarized how mass-spectrometry-based proteomics has matured from a largely technology-driven field of research into a mainstream analytical tool for the life sciences.

4. Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F *et al.*: **A human interactome in three quantitative dimensions organized by stoichiometries and abundances**. *Cell* 2015, **163**:712-723.

5. Sharma K, D'Souza RC, Tyanova S, Schaab C, Wisniewski JR, Cox J, Mann M: **Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling**. *Cell Rep* 2014, **8**:1583-1594.

6. Itzhak DN, Tyanova S, Cox J, Borner GH: **Global, quantitative and dynamic mapping of protein subcellular localization**. *Elife* 2016:5.

7. Robles MS, Cox J, Mann M: **In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism**. *PLoS Genet* 2014, **10**:e1004047.

8. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M: **Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ**. *Mol Cell Proteomics* 2014, **13**:2513-2526.

9. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ: **The one hour yeast proteome**. *Mol Cell Proteomics* 2014, **13**:339-347.

10. de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M: **Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast**. *Nature* 2008, **455**:1251-1254.

11. Stefely JA, Kwiecien NW, Freiberger EC, Richards AL, Jochem A, Rush MJ, Ulbrich A, Robinson KP, Hutchins PD, Veling MT *et al.*: **Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling**. *Nat Biotechnol* 2016.

12. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, Mann M: **Deep proteome and transcriptome mapping of a human cancer cell line**. *Mol Syst Biol* 2011, **7**:548.

13. Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J, Aebersold R: **The quantitative proteome of a human cell line**. *Mol Syst Biol* 2011, **7**:549.

14. Munoz J, Low TY, Kok YJ, Chin A, Frese CK, Ding V, Choo A, Heck AJ: **The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells**. *Mol Syst Biol* 2011, **7**:550.

15. Mann M, Kulak NA, Nagaraj N, Cox J: **The coming age of complete, accurate, and ubiquitous proteomes**. *Mol Cell* 2013, **49**:583-590.

16. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *J Am Soc Mass Spectrom* 1994, **5**:976-989.

17. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**:3551-3567.

18. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm.** *J Proteome Res* 2004, **3**:958-964.

19. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra**. *Bioinformatics* 2004, **20**:1466-1467.

20. Bern M, Cai Y, Goldberg D: **Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry**. *Anal Chem* 2007, **79**:1393-1400.

21. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R: **The Orbitrap: a new mass spectrometer**. *J Mass Spectrom* 2005, **40**:430-443.

22. Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M: **Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap**. *Mol Cell Proteomics* 2005, **4**:2010-2021.

23. Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M, Horning S: **Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer**. *Mol Cell Proteomics* 2011, **10** M111.011015.

24. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J: **The perseus computational platform for comprehensive analysis of (prote)omics data**. *Nat Methods* 2016, **13**:731-740.

25. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**. *Nat Biotechnol* 2008, **26**:1367-1372.

26. • Tyanova S, Temu T, Cox J: **The MaxQuant computational platform for mass spectrometry-based shotgun proteomics**. *Nat Protoc* 2016, **11**:2301-2319.
This is a protocol describing the usage of MaxQuant for shotgun proteomics data analysis on a large variety of experimental designs and quantification strategies.

27. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M: **Andromeda: a peptide search engine integrated into the MaxQuant environment**. *J Proteome Res* 2011, **10**:1794-1805.

28. Cox J, Michalski A, Mann M: **Software lock mass by two-dimensional minimization of peptide mass errors**. *J Am Soc Mass Spectrom* 2011, **22**:1373-1380.

29. Cox J, Mann M: **Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap**. *J Am Soc Mass Spectrom* 2009, **20**:1477-1485.

30. Patti GJ, Yanes O, Siuzdak G: **Innovation: metabolomics: the apogee of the omics trilogy**. *Nat Rev Mol Cell Biol* 2012, **13**:263-269.

31. Fuhrer T, Zamboni N: **High-throughput discovery metabolomics**. *Curr Opin Biotechnol* 2015, **31**:73-78.

32. Zhang A, Sun H, Wang P, Han Y, Wang X: **Modern analytical techniques in metabolomics analysis**. *Analyst* 2012, **137**:293-300.

33. Misra BB, van der Hooft JJ: **Updates in metabolomics tools and resources: 2014–2015**. *Electrophoresis* 2016, **37**:86-110.

34. Alonso A, Marsal S, Julia A: **Analytical methods in untargeted metabolomics: state of the art in 2015**. *Front Bioeng Biotechnol* 2015, **3**:23.

35. •• Cho K, Mahieu NG, Johnson SL, Patti GJ: **After the feature presentation: technologies bridging untargeted metabolomics and biology**. *Curr Opin Biotechnol* 2014, **28**:143-148.
In this publication emerging technologies that can be applied after untargeted profiling to extend biological interpretation of metabolomic data are reviewed. Recent advances are highlighted that help transform untargeted profiling results into structures, concentrations, pathway fluxes and localization patterns.

36. • Sevin DC, Kuehne A, Zamboni N, Sauer U: **Biological insights through nontargeted metabolomics**. *Curr Opin Biotechnol* 2015, **34**:1-8.
The authors compare the contributions of traditional targeted and non-targeted metabolomics in advancing different research areas. They conclude that novel computational approaches are required to tap the full potential of nontargeted metabolomics.

37. Michalski A, Cox J, Mann M: **More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC–MS/MS**. *J Proteome Res* 2011, **10**:1785-1793.

38. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical entities of biological interest: an update**. *Nucleic Acids Res* 2010, **38**:D249-D254.

39. Kind T, Fiehn O: **Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm**. *BMC Bioinform* 2006, **7**:234.

40. Zamboni N: **^{13}C metabolic flux analysis in complex systems**. *Curr Opin Biotechnol* 2011, **22**:103-108.

41. Munger J, Bennett BD, Parikh A, Feng XJ, McArdle J, Rabitz HA, Shenk T, Rabinowitz JD: **Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy**. *Nat Biotechnol* 2008, **26**:1179-1186.

42. Yuan J, Bennett BD, Rabinowitz JD: **Kinetic flux profiling for quantitation of cellular metabolic fluxes**. *Nat Protoc* 2008, **3**:1328-1340.

43. Wiechert W, Noh K: **Isotopically non-stationary metabolic flux analysis: complex yet highly informative**. *Curr Opin Biotechnol* 2013, **24**:979-986.

1.4 Multi-Omics Data Analysis

Arguably, the most challenging task is to take what are in essence snapshots of the state of a biological system under study, and combine them in a manner from which meaningful information could be extracted (Subramanian *et al.*, 2020). There are many reasons for studying more than one omics dataset simultaneously. The main challenge in such efforts, besides the handling of the huge amounts of data generated, is the heterogeneous and multidimensional nature of the data since each omics dimension is measured using a different technology, and is presented in a unique manner (Kim and Tagkopoulos, 2018). These differences can arise from the nature of the data, e.g. being discrete or continuous, or from the complexity of the measured omics dimension, e.g. the expression levels of tens of thousands of mRNAs, and the levels of thousands of metabolites, not to mention the differing sensitivity and reproducibility of each different technology. On the other hand, multi-omics efforts can be helpful in reducing noise and false positive findings within each omics dimension by aggregating data and evidence from several different layers of information (Rotroff and Motsinger-Reif, 2016).

Cross analysis of proteomics data with genomics data can correlate personalized hereditary or disease-related information to proteomic phenomena, such as correlating DNA copy number and loss of heterozygosity to protein expression by grouping together proteins matching to the same gene (Geiger, Cox and Mann, 2010a). By comparing the transcriptome to the proteome, the dynamic phenomena of gene expression between transcription and translation becomes detectable. When combining proteomics and transcriptomics data, expression levels may be the easiest to integrate due to their near 1:1 relationship (Tyanova *et al.*, 2016). An even closer relationship exists between transcriptomics and proteomics when one considers data from techniques such as ribosome profiling (Ingolia, 2014). Metabolomics data combined with proteomics data bears the possibility to study the interplay of enzymes with reaction reactants, since metabolites and proteins have an organic connection via metabolic reactions where enzymes are responsible for the consumption and production of metabolites. Metabolites can also act as catalysts, allosteric regulators and help form protein complexes (Piazza *et al.*, 2018).

1.4.1 Network Assisted Data Analysis

When analyzing data from different levels of omics, existing knowledge regarding established relationships between various biomolecules ranging from DNA to metabolites promises a logical approach to integrating such data (Y. X. Chen *et al.*, 2020). Many public databases are now available where data is curated by mining literature, where various interactions are intuitively represented as networks (Orchard *et al.*, 2014; Szklarczyk *et al.*, 2019; Oughtred *et al.*, 2020). Metabolic networks reconstructed on the level of various organisms provide a great opportunity to integrate various omics data and analyze them together (Büchel *et al.*, 2013). This is because the metabolome is the closest level to the phenotype of the organism of interest. The metabolic network provides the scaffold upon which all omics data can be mapped for integrative analysis (Chong and Xia, 2017). Such an approach in utilized in the paper presented in section 4.3 and discussed in detail.

2. Purpose

Within the context of this thesis, several project have been carried out in aims of improving computational methods for LC-MS/MS-based proteomics, metabolomics, and downstream analysis of multi-omics datasets. To this end, a new algorithm is proposed for improved mass accuracy in LC-MS/MS-based metabolomics datasets, which incorporates a novel library-based mass recalibration approach. This will in turn help increase the number of identifications in future metabolomics software and help propel metabolomics to the level of maturity that proteomics has reached. Furthermore, MaxQuant 2.0 equipped with MaxDIA is described for analyzing DIA LC-MS/MS proteomics datasets, using both measured libraries and predicted libraries. This further expands the abilities of the MaxQuant platform as the go to platform for the quantitative analysis of proteomics datasets. Finally, the Metis plugin for the Perseus software is introduced as an easy and accessible tool for metabolic network-based multi-omics analysis. Metis expands the capabilities of the popular Perseus software in network-based analysis and handling of various types of omics data.

3. Results

Here, the primary results on the improvement of mass accuracy in LC-MS/MS-based metabolomics data is presented. Later in chapter 4, the relevant publications to the improvements of the MaxQuant software suit in terms of supporting the Linux operating system and DIA proteomics data is presented, along with the Metis plugin for multi-omics data analysis within the Perseus software suit.

3.1 Metabolomics Library Generation

In order to generate a the initial library to use for mass recalibration in metabolomics, *m/z* values which are known common metabolites in biological samples from databases such as ChEBI (de Matos *et al.*, 2010), which accumulate and curate metabolites of biological importance are gathered. It is important to mention that the library is made up of all possible isotopic peaks, rather than simply monoisotopic masses for the metabolites, since features are to be matched to the library rather than isotope patterns. Subsequently, publically available metabolomics data from several resources, including MetaboLights (Haug *et al.*, 2020) and Metabolomics Workbench, were gathered and filtered for data from Orbitrap mass spectrometers, due to the higher resolution that this type of mass spectrometer provides for the lower mass range. This resulted in 71 datasets corresponding to 1511 runs. All data were then processed using the MaxQuant software for feature detection and using our novel mass morphing algorithm, features are mapped to the library. The library can then be updated if there remains unmapped features with adequate signal-to-noise ratio with a plausible new metabolite annotation (Figure 3.1).

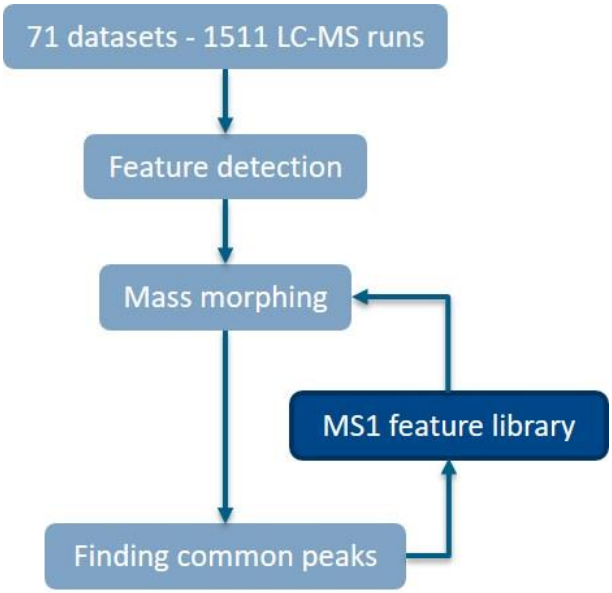


Figure 3.1: Metabolite library generation workflow.

Following the library generation and update strategy, the size of the library increases as newly identified high quality features within each dataset that are not present within the library of plausible *m/z* values are added to the library, making the library more comprehensive. This behavior plateaus as the library is updated with each iteration of processing the same dataset and will continue with every new dataset (Figure 3.2).

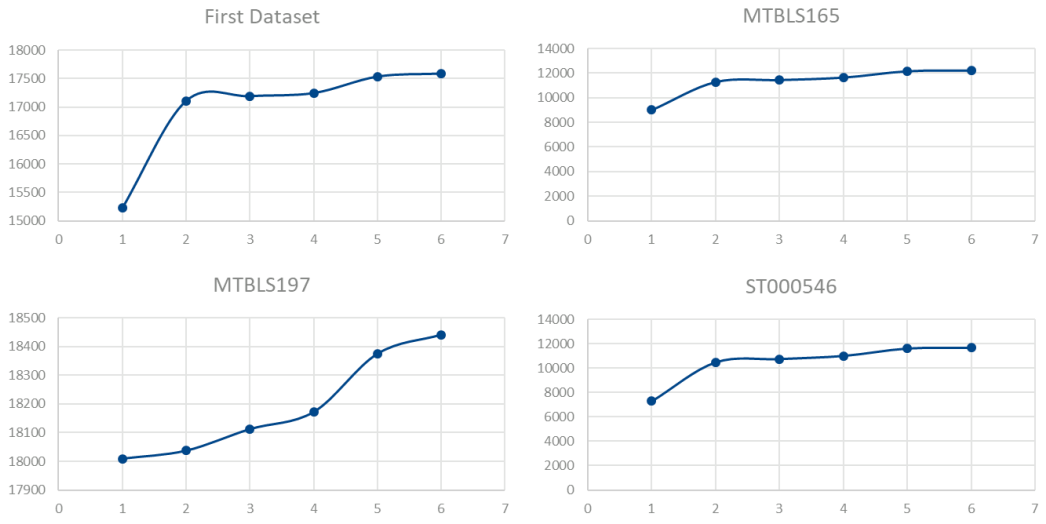


Figure 3.2: Library size increases with subsequent mapping and mass morphing and plateaus after several iterations on four datasets. The y-axis is the number of m/z vales present within the library and the x-axis is the number of iterations of library mapping and update.

The number of identified features within each dataset decreases as the growth of the library leads to higher mass accuracy and thus, more confident identifications. This is not the case for the initial dataset used to develop the algorithm and the initial library as many of the features within that dataset were initially manually identified (Figure 3.3).

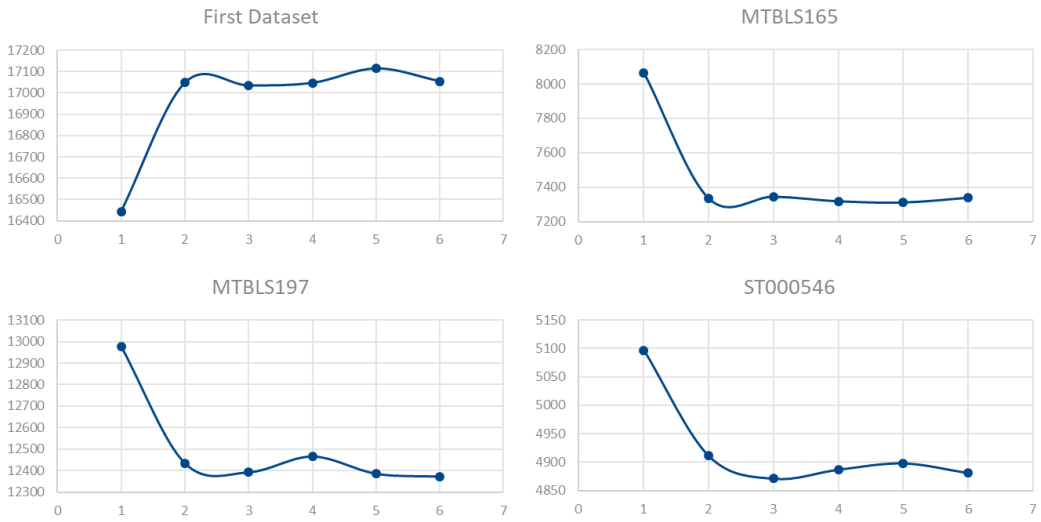


Figure 3.3: Number of identified features within four different mass spectrometry runs. The y-axis is the number of identified and the x-axis is the number of iterations of mass recalibration.

3.2 Library mapping, mass morphing and recalibration

We use our Easy Library Implementation (ELI) for mass recalibration in metabolomics, for a significant improvement in mass accuracy of metabolomics datasets. After generation of the library of plausible *m/z* values, the mass spectrometry run to be analyzed is processed and aligned to the library using our mass morphing approach, which calculates a nonlinear calibration function aiming to map as many features to the library as possible. Special attention is made to the smoothness of the fit to prevent overfitting (Figure 3.4).

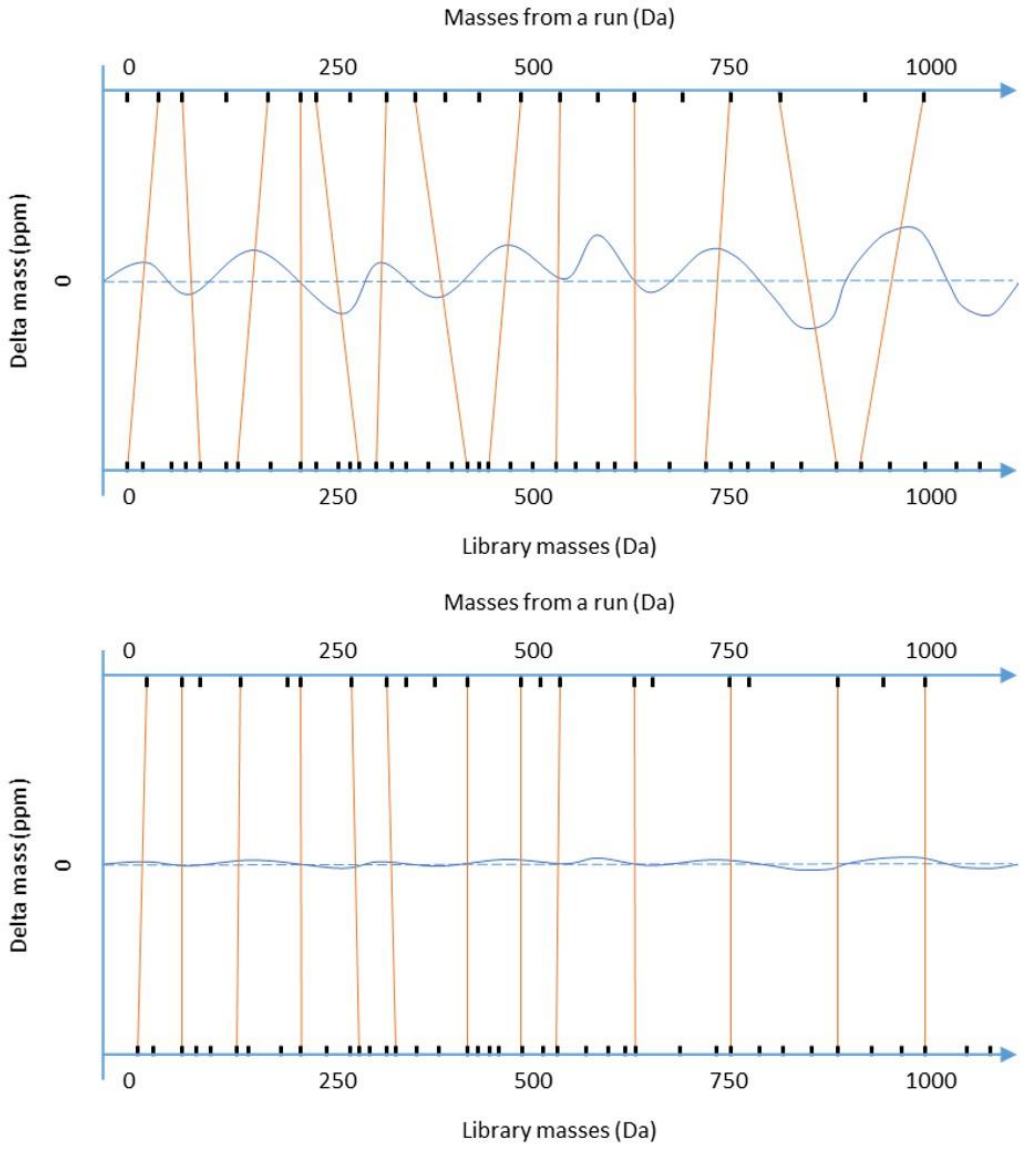


Figure 3.4: Schema of how features are mapped and morphed to the library.

ELI for mass recalibration in metabolomics when performed on all 1511 metabolomics runs that were gathered results in a general improvement across all datasets with the median mass error being reduced, with significant improvements in datasets which were suffering from high rates of mass error (Figure 3.5). Although mass accuracy improvements depends on factors such as molecular mass and the complexity of the sample, improvements in mass accuracy generally lead to the reduction of the number of candidates for each feature and thus, help in the effective processing analysis of metabolomics datasets.

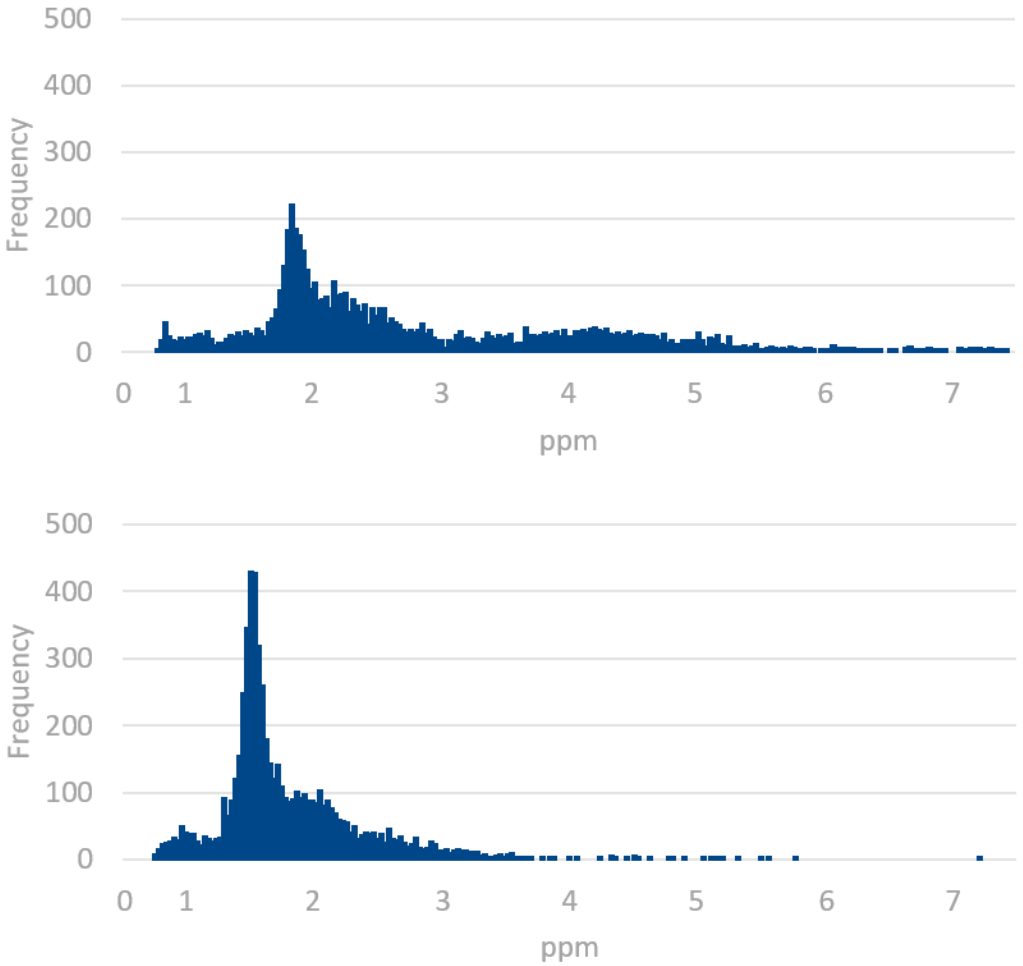


Figure 3.5: Average FWHM of un-calibrated vs. calibrated delta ppm across 1511 metabolomics runs. The mass error is significantly reduced in datasets that suffered from mass errors higher than 3 ppm and the median has reduced to be below 2 ppm.

4. Manuscripts

During the past few years, we have published papers on improvements to the MaxQuant software suit, multi-omics capabilities of the Perseus software suit and the introduction of MaxQuant 2.0, which will be presented in the following sections.

4.1 MaxQuant goes Linux

MaxQuant has been accepted by the proteomics community as the gold standard in analyzing proteomics data for more than a decade. However, due to its Windows only structure and the limitations of Windows in running powerful servers with many hundreds of CPU cores, many larger proteomics projects suffered from lengthy run times. Adapting the MaxQuant code base to Linux-based operating systems has not only allowed larger proteomics projects to be processed on larger servers running Linus-based operating systems, it has also allowed the more advanced MaxQuant user to utilize MaxQuant in custom scripts and workflows for streamlined analysis. I have been privileged to be part of the team involved in this development in the MaxQuant ecosystem by contributing to the transition to Linux and testing its performance.

Contributions to the following correspondence within the context of this thesis include software development and research into cross platform software development strategies.

Sinitcyn, Pavel, Shivani Tiwary, Jan Rudolph, Petra Gutenbrunner, Christoph Wichmann, Şule Yllmaz, **Hamid Hamzeiy**, Favio Salinas, and Jürgen Cox. 2018. “MaxQuant Goes Linux.” Nature Methods 15 (6): 401. https://doi.org/10.1038/s41592-018-0018-y.

correspondence

MaxQuant goes Linux

To the Editor: We report a Linux version of MaxQuant[1] (http://www.biochem.mpg.de/5111795/maxquant), our popular software platform for the analysis of shotgun proteomics data.

One of our main intentions in developing MaxQuant was to 'take the pain out of' quantifying large collections of protein profiles[2]. However, unlike, for instance, the Trans-Proteomic Pipeline[3], the original version of MaxQuant could be run only on Microsoft Windows, and thus its use was restricted in high-performance computing environments, which very rarely use Windows as an operating system. When we began developing MaxQuant, Windows was the only operating system supported by vendor-provided raw data access libraries. Therefore, we wrote MaxQuant in the C# programming language on top of the Windows-only .NET framework. Windows support for cloud platforms is more expensive, and the operating system is harder to use and less scalable compared with Linux.

We recently carried out a major restructuring of the MaxQuant codebase, and we made it compatible with Mono (https://www.mono-project.com/), an alternative cross-platform implementation of the .NET framework. Furthermore, we now provide an entry point to MaxQuant from the command line without the need to start its graphical user interface, which allows execution from scripts or other processing tools. Meanwhile, Thermo Fisher Scientific has released its platform-independent and Mono-compatible implementation of its raw data access library (http://planetorbitrap.com/rawfilereader), and hopefully more vendors will follow soon. Together, this leads to a situation in which large-scale computing of proteomics data with MaxQuant becomes feasible on all common platforms.

When we parallelized the MaxQuant workflow over only a few central processing unit (CPU) cores, we hardly noticed a difference in performance between Linux and Windows (Fig. 1). However, in benchmarking of a highly parallelized MaxQuant run on 120 logical cores, we observed that the Linux version showed highly superior parallelization performance, with speed 64% faster than that observed under a Windows server operating system using identical hardware. MaxQuant uses operating system processes, rather than the intrinsic multi-threading mechanism of C#, to realize parallel execution, and it manages the load-balancing of an arbitrarily large set of raw data files over a specified number of processors by itself. We hypothesize that this allows Linux to optimize parallel execution to the high extent that we observed. A larger benchmark study is under way, in which we will investigate the dependence of the increased speed on hardware such as, for instance, the type of CPU and storage systems.

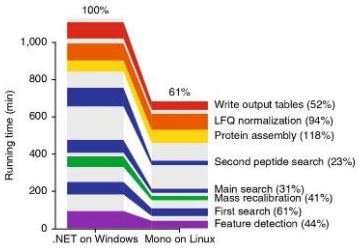


Fig. 1 | Benchmarking MaxQuant on Linux and Windows. We analyzed 300 LC-MS runs with MaxQuant using 120 logical cores in parallel, once with Ubuntu Linux (version 16.04.3) and once with Windows server 2012 R2 as the operating system. We used identical hardware in both cases: four Intel Xeon E7-4870 CPUs and 256 GB of DDR3 RAM. The total running times are shown, and several long-running sub-workflows are highlighted. Percentages indicate the amount of time needed to complete the relevant process in Linux as a percentage of the total time required for the same process in Windows.

MaxQuant has already been adapted in several forms for cloud and high-performance computing applications, as described, for instance, by Judson et al.[4] and on the Chorus platform (https://chorusproject.org). We expect that the number of applications will increase with our Linux-compatible MaxQuant version. We envision that proteomics core facilities, for instance, will benefit from the combination of command-line access and Linux compatibility, which enables standardized high-throughput data analysis. The MaxQuant code base is identical for Windows and for Linux; thus there is only a single distributable running on both operating systems, which can be downloaded from http://www.maxquant.org (version 1.6.1.0). MaxQuant is freeware, and contributions to new functionality are collaboration-based. The code of open source parts is available at https://github.com/JurgenCox/compbio-base. ❐

Pavel Sinitcyn, Shivani Tiwary, Jan Rudolph, Petra Gutenbrunner, Christoph Wichmann, Şule Yılmaz, Hamid Hamzeiy, Favio Salinas and Jürgen Cox*

Computational Systems Biochemistry, Max Planck Institute for Biochemistry, Martinsried, Germany.

**e-mail: cox@biochem.mpg.de*

Published online: 31 May 2018
https://doi.org/10.1038/s41592-018-0018-y

References

1. Cox, J. & Mann, M. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
2. Azvolinsky, A., DeFrancesco, L., Waltz, E. & Webb, S. *Nat. Biotechnol.* **34**, 256–261 (2016).
3. Deutsch, E. W. et al. *Proteomics Clin. Appl.* **9**, 745–754 (2015).
4. Judson, B., McGrath, G., Peuchen, E. H., Champion, M. M. & Brenner, P. In *Proc. 8th Workshop on Scientific Cloud Computing* (eds. Chard, K. et al.) 17–24 (ACM, New York, 2017).

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program (grant agreement no. 686547 to J.C., J.R. and S.Y.) and from the FP7 (grant GA ERC-2012-SyG_318987–ToPAG to S.T. and F.S.).

Author contributions

P.S., S.T., J.R., P.G., C.W., S.Y., H.H., F.S. and J.C. developed the software. P.S. conducted the performance analysis. J.C. wrote the manuscript.

Competing interests

The authors declare no competing interests.

4.2 MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics

MaxQuant 2.0 follows in the lineage of the original MaxQuant software published more than a decade ago, which has become the gold standard go to software for the processing of raw LC-MS/MS data. Originally, MaxQuant was designed and implemented for DDA experiments, and now, analysis of data-independent acquisition data can be carried out by the MaxDIA algorithm. In the context of this PhD work, all relevant testing data was gathered, benchmarking was performed, machine-learning algorithms within the workflow were optimized, external collaborations were coordinated, and the following manuscript was written along with the other co-authors.

Contributions to the following correspondence within the context of this thesis include software design and development, software benchmarking, data analysis and ensuring support for external resources such as the PRIDE repository.

Pavel Sinitcyn, **Hamid Hamzeiy**, Favio Salinas Soto, Daniel Itzhak, Frank McCarthy, Christoph Wichmann, Martin Steger, Uli Ohmayer, Ute Distler, Stephanie Kaspar-Schoenefeld, Nikita Prianichnikov, Şule Yılmaz, Jan Daniel Rudolph, Stefan Tenzer, Yasset Perez-Riverol, Nagarjuna Nagaraj, Sean J. Humphrey and Jürgen Cox. “MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics.” Submitted to Nature Biotechnology, 2020

MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics

Pavel Sinitcyn[1,#], Hamid Hamzeiy[1,#], Favio Salinas Soto[1,#], Daniel Itzhak[2], Frank McCarthy[2], Christoph Wichmann[1], Martin Steger[3], Uli Ohmayer[3], Ute Distler[4], Stephanie Kaspar-Schoenefeld[5], Nikita Prianichnikov[1], Şule Yılmaz[1], Jan Daniel Rudolph[1,6], Stefan Tenzer[4], Yasset Perez-Riverol[7], Nagarjuna Nagaraj[5], Sean J. Humphrey[8] and Jürgen Cox[1,9,*]

[1]Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.

[2]Chan Zuckerberg Biohub, 499 Illinois St., San Francisco, CA 94158, USA.

[3]Evotec München GmbH, Am Klopferspitz 19a, 82152 Martinsried, Germany.

[4]Institute for Immunology, Johannes Gutenberg University, Langenbeckstraße 1, 55131 Mainz, Germany.

[5]Bruker Daltonik GmbH, Farenheitstr. 4, 28359 Bremen, Germany.

[6]Bosch Center for Artificial Intelligence, Robert-Bosch-Campus 1, 71272 Renningen, Germany

[7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD UK

[8]School of Life and Environmental Sciences, Charles Perkins Centre, University of Sydney, John Hopkins Drive, Camperdown, NSW 2006, Australia.

[9]Department of Biological and Medical Psychology, University of Bergen, Jonas Liesvei 91, 5009 Bergen, Norway.

[#]These authors contributed equally to the publication.

[*]Correspondence: cox@biochem.mpg.de

Abstract

MaxDIA is a universal platform for analyzing data-independent acquisition proteomics data within the MaxQuant software environment. Using spectral libraries, MaxDIA achieves cutting-edge proteome coverage with significantly better coefficients of variation in protein quantification than other software. MaxDIA is equipped with accurate false discovery rate estimates on both library-to-DIA match and protein levels, also when using whole-proteome predicted spectral libraries. This is the foundation of discovery DIA – a framework for the hypothesis-free analysis of DIA samples without library and with reliable FDR control. MaxDIA performs three- or four-dimensional feature detection of fragment data and scoring of matches is augmented by machine learning on the features of an identification. MaxDIA's novel bootstrap-DIA workflow performs multiple rounds of matching with increasing quality of recalibration and stringency of matching to the library. Combining MaxDIA with two new technologies, BoxCar acquisition and trapped ion mobility spectrometry, both lead to deep and accurate proteome quantification.

Data-independent acquisition (DIA) proteomics[1] promises robust and accurate quantification of proteins over large-scale study designs and across heterogeneous laboratory conditions[2]. In all omics sciences, robust data analysis pipelines are as important as the data acquisition technology itself, and proteomics is no exception. MaxQuant[3–6] is the most widely-used software for analyzing data-dependent acquisition (DDA) proteomics data, providing a vendor-neutral complete end-to-end solution for all common experimental designs. With version 2.0 described here, MaxQuant offers an equally complete DIA software infrastructure, termed MaxDIA. Such a unified framework over all mass spectrometry-based proteomics based on peptide quantification comes with several advantages over existing software[7–10]. DDA libraries and DIA samples can be processed in integrated, consistent ways. Algorithmic parts of the workflow that do not depend on the type of acquisition, like protein quantification algorithms, such as MaxLFQ[11], protein redundancy grouping, or protein-level false discovery rate (FDR) can be applied to all data in exactly the same way, making DDA and DIA studies much more comparable.

The classical approach to DIA data analysis utilizes a spectral library of peptides which are queried in the DIA samples and quantified in case of their presence. In this spectral library-based approach, the rate of false matches can in principle be controlled with techniques similar to those developed in DDA proteomics[12]. For instance, the target-decoy method[13] has been adapted to DIA[9]. Additionally, several library-free approaches exist[14] and spectral predictions have been successfully used for DIA data analysis[15–20]. However, effective control of false discovery rates, in particular on the level of identified proteins with these methods is still a critical aspect. Once this is achieved, DIA can additionally be employed in a discovery mode, without biases imposed by a library and with the certainty that the identified set of proteins contains at most a predefined percentage of false positives, e.g. 1%, as is standardly applied in DDA-based proteomics. Here we demonstrate that MaxDIA fulfills these criteria and can indeed be used in such a discovery DIA mode.

Machine learning is an integral part of MaxDIA. We use the bidirectional recurrent neural network[21] (BRNN) approach termed DeepMass:Prism[15] to create *in silico* very precise libraries of MS/MS spectra for peptides digested from complete proteome sequence databases. BRNNs are also used for the dataset-specific prediction of liquid chromatography retention times. Furthermore, to score library DIA sample matches based on multivariate information derived

from properties of the matches, we apply the gradient boosting method XGBoost[22], which is highly superior to only using the matching score itself, and also compared to applying other machine learning approaches.

High-quality three-dimensional (3D) or, in the presence of ion mobility data, 4D feature detection[3,23] of the precursor data is one of the most important ingredients of MaxQuant for DDA data, leading to efficient noise suppression. In MaxDIA, fragment ions are additionally detected as 3D/4D features. Besides noise removal, this ensures that data is not over-interpreted: The feature detection on fragment data allows to require that all signals belonging to a 3D/4D peak contribute as evidence only to one peptide identification, ensuring that signals at slightly different retention times or ion mobility values, but really belonging to the same feature, are not used as independent evidence for two similar peptides, e.g. differing by a modification or resulting from an amino acid polymorphism.

In MaxDIA we support two new and promising technologies, both of which enable deep quantification of DIA samples. One is to combine DIA with high dynamic-range precursor data obtained by the BoxCar acquisition method[24]. The second is to utilize ion mobility as an extra data dimension on a timsTOF Pro instrument[25–27] for DIA. Both increase the quantified proteome in DIA samples substantially providing highly precise and linear quantification over the whole dynamic range. Furthermore, since the MaxLFQ algorithm has been designed to perform label-free quantification on pre-fractionated samples[11], also MaxDIA has the capability to perform label-free quantification of pre-fractionated samples analyzed by DIA, which opens up applications of DIA requiring ultra-deep proteome quantification. Complete submissions to the PRoteomics IDEntifications[28] (PRIDE) database using an adapted mzTab[29] scheme can also be performed automatically using MaxDIA.

RESULTS

MaxDIA data analysis workflow

MaxDIA is embedded into the MaxQuant software environment (Fig. 1) and shares with it the graphical user interface, computational infrastructure, and many algorithmic workflow components applicable to both. It is vendor-neutral, with direct support for the most common native vendor file formats for reading mass spectra, as well as the open mzML file format[30].

MaxDIA can be operated in a classical library-based approach or in discovery DIA mode. In the former, DIA datasets are interrogated within MaxQunat by spectral libraries generated with MaxQuant, while the latter does not require acquisition of a spectral library. In discovery DIA mode, spectral libraries are generated by DeepMass:Prism[15], a bidirectional recurrent neural network that enables precise prediction of spectral intensities from peptide sequences. Decoy spectra are generated by reverting library sequences under the constraint of preserving the cleavage characteristics of the protease that was used in the experiment and ensuring that the decoy peptide masses, retention times and ion mobility values follow the same multivariate distribution as the target peptides. DIA samples and libraries are then analyzed in an end-to-end workflow for peptide and protein identification and quantification. MaxQuant's three-dimensional (3D) or four-dimensional (4D) feature detection[3,23] (Fig. 2) and de-isotoping is performed on the precursor data and on all LC-MS/MS or LC-IMS-MS/MS fragment data domains corresponding to precursor selection windows. Defining MS/MS features in a multi-dimensional way is particularly important for fragment data, since it avoids over-interpretation of identification results. This enables the requirement that every MS/MS feature is used at most once in peptide identification. Problems may arise if such precautions are not taken, since features will be double-counted for the identification of peptides that are similar to each other due to sequence homology or due to the presence or absence of a modification, but for which there is insufficient evidence for the existence of both peptide forms.

Bootstrap-DIA

Central to the workflow is bootstrap-DIA, which consists of multiple steps of matching the library spectra to DIA samples (Supplementary Fig. 1). These steps aim to bootstrap the DIA identification process based on the least possible prior knowledge. Bootstrap-DIA replaces and substantially extends the concept of the 'first search-main search' strategy[31] as well as the 'retention time alignment' and 'match between runs' used in DDA MaxQuant. Increasingly more information is gained in each round, with this information utilized in subsequent rounds. For instance, in the first round of matching, no retention time constraint is used. Based on these matches, a linear model is fit between the library and sample retention times, which is used to align runs to one another, even when gradient lengths substantially differ. This linear correction can be applied to the data and in the second round of matching, retention times can be filtered based on a time window that is automatically adapted to the distribution of all retention time differences after linear alignment. This filtering removes sufficiently many false positive

matches, so that from the third round of matching a nonlinear retention time recalibration function can be determined. Application of the nonlinear recalibration function allows to subsequently apply more stringent filtering. Similar multi-step recalibration and filtering steps are applied to precursor and fragment masses, as well as to collision cross sections, if applicable. Supplementary Fig. 2 shows how target decoy distributions are affected after each matching step with increasingly more stringent filers. The resulting nonlinear precursor and fragment m/z recalibrations depending on m/z and retention time are shown in Supplementary Figs. 3, 4.

A consequence of the bootstrap-DIA process is that precursor and fragment masses, retention times and ion mobility values are nonlinearly aligned between each DIA sample and library without the need for spike-in standards. A prerequisite for this is that the DDA runs in the datasets used for the library are well aligned to each other, since the precision of alignment between library and DIA samples is otherwise limited by the variability of retention times and collision cross sections within the library. Therefore, when processing libraries in MaxQuant, retention time and ion mobility alignments should be activated. A challenging attribute that can be learned from the data are nonlinear retention time mappings between library and samples. This means that gradients between library and DIA runs do not need to be the same, and label-free quantification is possible even between DIA measurements with different gradients lengths. To evaluate the matching of different DIA gradient durations to a library we generated a DDA library consisting of 16 high pH-reversed phase fractions of a HeLa cell lysate measured with 25-minute gradients, and measured the same sample unfractionated with DIA using 30, 60, 90 and 120-minute gradients. Supplementary Fig. 5 shows retention time alignments between the library and DIA samples, and precise quantification between samples with different gradient lengths are shown in Supplementary Fig. 6. These capabilities greatly enhance the flexibility of MaxDIA, making the software applicable to analyzing a broader range of samples.

Scoring of library-to-sample matches by machine learning

To quantify the quality of match between a library spectrum and a DIA sample at a given retention time (and CCS value) we first find a precursor feature and all fragment features that match to the library spectrum with tolerances for m/z, retention time and CCS, dependent on the matching step in the bootstrap-DIA workflow. To measure the match quality, we then

calculate a score which is the sum over all matching features of numbers between zero and one, each quantifying how far away from the apex the respective peak was hit (Supplementary Fig. 7). For a given library spectrum this score is maximized over retention time and ion mobility. It is then ensured, through a second round of scoring, that every feature in a DIA sample is used at most for one library spectrum match.

This score then is enhanced through machine learning. To this end, we construct a feature space that in addition to the score contains various properties of the match (Supplementary Fig. 8), such as mass errors (in p.p.m.) for precursor and fragments compared to masses calculated from elemental compositions, retention time and ion mobility errors, and apex fractions. We employ a classification algorithm to separate target from decoy hits based on this feature space. We define the machine learning based match score as the assignment probability to the target class of the machine learning algorithm. This is not just the binary decision of the classifier, but a number expressing the affinity to the target spectra as opposed to the decoy spectra. To eliminate the risk of overfitting, we determine these machine learning scores in 5-fold cross validation, such that a match for which the machine learning score is calculated has not been used for training the model that is used for its prediction.

We used several different classification algorithms and monitored their effect on the identification performance of MaxDIA. We compared the performances of XGBoost[22], fully connected multi-hidden layer neural networks, random forests[32] and AdaBoost (Supplementary Fig. 9) scanning for each algorithm suitable ranges of meta-parameters. We found that XGBoost performs best among the tested algorithms, in contrast to Demichev et al.[10] who found neural networks to perform favorably. This choice is also different from DDA where for similar purposes support vector machine based methods are used[33]. XGBoost provides information on the importance of features for classification (Supplementary Fig. 8). We found that in the library-based approach, the feature defining whether the precursor has an isotope pattern assigned or was only seen as a single peak is of greater importance than the raw score itself. Furthermore, retention time, precursor mass errors, number of modifications and missed cleavages were among the top 10 highest ranked features. Also among the top 10 is the 'sample fragment overlap' which quantifies if and to what extent the N- and C- terminal ion series are overlapping in the DIA sample, thereby placing restrictions on the precursor mass.

Identification performance and quantification precision

To evaluate the performance of MaxDIA we ran it, and Spectronaut 13 on a dataset comprising 27 technical replicate injections of peptides derived from the human HepG2 cell line measured in DIA as well as a DDA library created from 12 high pH-reversed phase fractions (see online Methods). Using default parameters in both software, including a 1% FDR on precursor and protein levels, we obtain 6,238 protein groups mapped to Entrez gene identifiers with MaxDIA, compared to 6,015 with Spectronaut with an overlap of 5,549 (Fig. 3a). MaxDIA finds 20% more peptides than Spectronaut at 1% library-to-DIA matches FDR. The length distribution of identified peptides is very similar between the two analysis software (Fig. 3b).

While DIA is believed to be better in terms of data completeness[34,35] compared to DDA, we observe that this depends on the algorithmic details and that there is a tradeoff between data completeness and confidence of protein identification within a specific sample, as opposed to the whole dataset. After identifying peptides and proteins for the whole dataset, we apply a 'transfer q-value' cutoff to the identifications of matches in each sample. Setting it to 1, implies that no sample-specific restrictions are applied and that the peptide is quantified, whenever any evidence is found for its existence. A transfer q-value of 0.01 (equal to the global q-value of library-to-sample matches) results in stringent identification in every sample and hence, certainty about the actual sample-specific presence of peptides and proteins. We scanned through 7 values of the transfer q-value between 0.01 and 1 and monitored the number of proteins which have a certain number or less valid values in terms of LFQ intensities (Fig. 3c). As expected, for larger transfer q-values, the curves are flatter and higher in terms of total protein numbers. When using 1 for the 'minimum ratio count' parameter of the LFQ algorithm, most parts of all curves are above the line for the Spectronaut software. For 'minimum ratio count' = 2, which ensures higher accuracy of quantification, the array of curves is intersecting with the Spectronaut curve. After evaluating the accuracy of benchmark quantification results on several mass spectrometry platforms we decided to select 0.3 as the default value for the transfer q-value. Study-specific objectives (completeness of quantification vs. certainty of identification in individual samples) may suggest deviations from this default value.

The distribution of coefficients of variation (CVs) (Fig. 3d) indicates substantially higher quantification precision obtained with MaxLFQ (described below) in MaxDIA compared with Spectronaut, with median CVs of 0.072 and 0.109, respectively. Fig. 3e.f show typical log-log

scatter plots of protein intensities between replicates displaying less outliers and higher Pearson correlation for MaxDIA. All pair-wise replicate Pearson correlations of logarithmic intensities are represented as a heat map in Fig. 3g for both programs, showing consistently higher correlations for MaxDIA (median 0.993) compared to Spectronaut (median 0.977). We find a good overall agreement between averaged Spectronaut intensities and MaxDIA iBAQ values (Fig. 3h) with a Pearson correlation of 0.87. We performed mRNA vs. protein copy number comparisons based on RPKM[36] and iBAQ[37] values, respectively, using MaxDIA and Spectronaut (Fig. 3i,j). Both comparisons show similar correlations between mRNA and protein levels, which are also compatible with correlations typically found in such studies[38].

Accuracy of FDR estimates and discovery DIA

In order to evaluate the reliability of FDR estimates using MaxDIA's target-decoy strategy, we used a pooled DDA library generated from mixed human and maize samples, with corresponding DIA runs comprising only human samples[34]. Hence, every match identified as being derived from the maize proteome is a known false positive identification (having discarded peptides that are shared between proteins of the two species). This enables calculation of an 'external' FDR which is calculated independently of the 'internal' FDR estimated by the decoy approach in MaxDIA. Fig. 4a compares internal and external FDRs on match, peptide and protein group levels. The curves for internal and external FDR are in very good agreement on all three levels. When comparing the numbers of identified matches, peptides and protein groups at 1% FDR, which is often taken as a default value in shotgun proteomics, the numbers differed only by 3.0%, 3.4% and 5.0%, respectively, between internally and externally controlled FDR. Hence our decoy-based FDR estimates are in good agreement with external FDR calculations.

Given these results, we investigated how accurate the FDR estimates are for cases in which the library is dissimilar to the DIA sample. Hence, we assembled a library of *in-silico* predicted spectra based on DeepMass:Prism[15] consisting of all tryptic peptides digested from all human UniProt[39] sequences (Release 2019_05 containing 20959 proteins) without missed cleavages. We additionally generated predicted retention times for each *in-silico* spectrum based on a bidirectional recurrent neural network used previously for the same purpose[15]. Using this library with the same DIA dataset as in Fig 4a, we generated the same curves for internal and external FDRs as before (Fig. 4b). Also here we observed good agreement between internal

and external FDRs. In particular, at an FDR of 1% the number of identified protein groups differed only by 1.5%. We do however identify 39% more protein groups with the *in-silico* library compared to with the measured library. This highlights that MaxDIA does not require that spectral libraries are generated from matching samples in a project-specific manner, and yet FDRs are still reliably controlled. This enables the use of MaxDIA in a 'discovery' mode (discovery DIA), which is not biased by a library and completely hypothesis-free in terms of which proteins can be found, by using *in-silico* predicted libraries for all protein sequences.

We additionally repeated these analyses using the raw matching score instead of the machine learning-improved score (Fig 4c, d). This revealed that the agreement of internal and external FDR does not depend on whether the XGBoost-based machine learning was used to adjust the scoring. However, the use of machine learning does substantially increase peptide (83% and 58% for library DIA and discovery DIA, respectively) and protein group identifications (28% and 18%, respectively).

MaxLFQ adaptation for DIA

A prime example of the re-use and continued development of algorithms from DDA MaxQuant to MaxDIA is the label-free quantification algorithm, MaxLFQ[11]. Here, quantification is based on first calculating all pair-wise peptide ratios between samples, which are then summarized by the intensity profile that best fits all the pair-wise ratios. This procedure can be generalized to DIA by replacing a single ratio per peptide with multiple ratios derived from precursor intensities and from the most intense fragment peaks (Supplementary Fig. 10). This approach naturally implements hybrid quantification of precursor and fragment intensities.

To benchmark quantification accuracy, we downloaded a four-species dataset with well-defined small ratios between replicate groups[34]. Ratios are expected to be 0%, 10%, 20% or 30%, depending on the species comprising: *H. sapiens*, *C. elegans*, *S. cerevisiae* and *E. coli*. We tested several combinations of precursor, fragment or mixed quantification and fragment intensities summed up or kept separately. We measured the variability as the inter-quartile range of ratios within each species, and summed these over the four species (Fig. 5a). We found that hybrid quantification between precursors and fragments with fragment intensities kept separate for individual ion types in LFQ resulted in the smallest quantification errors measured as the sum of the inter-quartile ranges of ratio distributions over the four species. The accuracy

observed exceeded both MS1- and MS2-level quantification reported by Bruderer et al.[34]. A further question is how the filtering of fragments by their intensity improves quantification accuracy. To this end, we used only the top-N intense peaks for quantification while varying N (Supplementary Fig. 11a). We found that accuracy increases with the number of fragments used, indicating that no filtering of fragments by intensity is required. Similarly, we investigated, if filtering to top-N most intense peptides per protein is beneficial (Supplementary Fig. 11b), finding that it is best to use all available peptides.

Next, we analyzed a quantitative benchmark dataset obtained on SCIEX TripleTOF 6600 instrument, mixing proteomes from three species in defined ratios between replicate groups[2] (Fig. 5b). Using the original library analyzed with MaxQuant and using default values for all parameters, we identify 4,627 protein groups and achieve linear quantification for all three species over the whole dynamic range. In discovery mode with a predicted library allowing for one missed tryptic cleavage, the number of identified protein groups raises by 48% to 6,858 (Fig 5c) with on average improved quantification accuracy for the species with ratios as measured by inter-quartile ranges of species-specific ratio distributions. Importantly, *H. sapiens* which expresses a much larger number of proteins received the largest increase, identifying almost 2-fold more protein groups (4,012 vs. 2,127), while *C. elegans* and *E. coli* received proportionally fewer additional proteins.

We next acquired a quantitative three-species benchmark dataset utilizing ion mobility on a Bruker timsTOF Pro instrument. Using the DDA library acquired on the same instrument type, we identify 10,352 protein groups. We again used MaxLFQ for DIA with hybrid quantification with separate intensities for each fragment ion (Fig. 5d), seeing excellent quantification over the whole dynamic range without nonlinearities. In discovery mode (Fig. 5e), the number of identified protein groups increases to 10,466 with higher quantification accuracy, again judged by the inter-quartile ranges of ratio distributions. Scanning through the transfer q-value, we found that quantification accuracy was best with a value near 0.3 (Supplementary Fig. 12).

BoxCar and fractionated DIA

We recently implemented analysis of data acquired using the BoxCar acquisition method in MaxQuant in the DDA context[24], whose primary goal is to achieve higher dynamic range for the precursor intensities. Since this should be beneficial for DIA as well, we implemented its

generalization to combining high-dynamic range precursor measurements with DIA acquisition for the fragments. Furthermore, it is possible with MaxDIA to analyze and quantify DIA samples that have been pre-fractionated on peptide or protein levels. To showcase these features, we acquired both DDA libraries and DIA measurements from HEK cell lysate as single shots and as high-pH reversed phase peptide fractionated samples, which were pooled into eight fractions for MS analysis (see Online Methods). We analyzed all combinations of libraries and samples, and in addition we analyzed the DIA samples in discovery-DIA mode allowing for one missed trypsin cleavage (Fig. 6a). For the fractionated DIA samples we observe an increase in the number of identified protein groups concomitant with the size of the library, with the most identifications in discovery mode. With single shot samples, the number of identified proteins saturates with library size, having slightly more identifications with the fractionated library. However, comparing identifications for the single shot DIA samples between fractionated library and discovery mode, we find that the results are very similar with 89% overlap of Entrez gene identifier mapped protein groups (Supplementary Fig. 13). This indicates that for both types of DIA samples it is not compulsory to produce a deep, fractionated library, but that comparable or even better results can be achieved in discovery DIA mode. Quantification with MaxLFQ between three replicates of fractionated DIA samples shows very good correlation with a median Pearson correlation of 0.993 (Fig. 6b).

We then compared the results obtained with the three different library-creation approaches to RNA-seq data of HEK cells (see Online Methods). Fig. 6c compares the four sets of identifications based on gene identifiers. Out of the 9,503 genes covered by proteomics methods, 65% were found with all three library methods. Additional 25% were found with both, discovery mode and fractionated library, but not with the single shot library. 608 proteins were uniquely found with the discovery approach, compared to 251 with the deep fractionated library, suggesting preference for the discovery mode from the perspective of results, in addition to its economic advantages. In Fig. 6d, the results from Fig. 6c are displayed according to RPKM intervals of the RNA-seq data. The RNA-seq data shows a bimodal left shoulder that is typical of expression noise[40], genes for which there is only limited proteomic evidence of translation. As expected, highly abundant proteins are recovered with all methods, while at low abundance, both the deep-fractionated library and discovery DIA approach add identifications.

DISCUSSION

Here we introduce MaxDIA, a complete end-to-end DIA workflow embedded into the MaxQuant environment with major new features and broad applicability to established and novel mass spectrometry technologies. We demonstrate the widespread and general utility of the software, including its use in analyzing BoxCar-DIA and ion mobility DIA data, demonstrating very high proteome quantification coverage.

This framework lends itself to several extensions which are currently under development. In particular, while the analysis of posttranslational modifications (PTMs) is possible in principle by providing suitable libraries with spectra from modified peptides, proper localization of the modification on the peptide has to be carefully implemented as an additional process following peptide identification[41]. For these purposes, a PTM score guiding localization needs to be calculated directly from the DIA data and not from extracted spectra. Similarly, extensions to the identification of cross-linked peptides are straightforward[42] and are planned for future releases of MaxDIA.

ONLINE METHODS

HepG2 technical replicate data

Cell culture and MS sample preparation. HepG2 were from ATCC and cultured in MEM and 10% FCS. Cells were washed twice with ice-cold PBS and harvested using freshly prepared SDC buffer (1% SDC, 10 mM TCEP, 40 mM CAA, 75 mM Tris-HCl at pH= 8.5). The SDC lysates were heated to 95°C for 10 min while shaking at 750 rpm in a Thermomixer (Eppendorf) and then sonicated for 10 min (10 x 30 sec on/off cycles) using a Bioruptor® Pico sonication device (Diagenode). Protein concentrations were determined using the 660 nm assay (Thermo Fisher Scientific) and the proteins were digested with trypsin/Lys-C mix (Promega, V5071) overnight at 37°C with a 1:50 enzyme to protein ratio. The digestion was stopped by adding two volumes of 99% ethylacetate/1% TFA, followed by sonication for 1 min using an ultrasonic probe device (energy output ~40%). The samples were then desalted using in-house prepared, 200 µl two plug SDB-RPS StageTips[43] (3M EMPORETM, 2241). SDB-RPS StageTips were conditioned with 60 µl isopropanol, 60 µl 80% ACN/5% NH4OH and 100 µl 0.2% TFA. The SDC/ethylacetate mixture was directly loaded onto the tips followed by two washing steps of 200 µl 0.2% TFA each. Peptides were eluted with 80% ACN/5% NH4OH, speedvac dried and then resupended in 0.1% FA. After estimation of the concentration using a nanodropTM device (Thermo Fisher Scientific), the samples were adjusted to 0.4 µg/µl with 0.1% FA, of which 2 µl (800 ng) were injected into the mass spectrometer.

LC-MS/MS measurements. Peptides were loaded on 40 cm reversed phase columns (75 µm inner diameter, packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin [ReproSil-Pur®, Dr. Maisch GmbH]). The column temperature was maintained at 60°C using a column oven. An EASY-nLC 1200 system (ThermoFisher) was directly coupled online with the mass spectrometer (Q Exactive HF-X, ThermoFisher) via a nano-electrospray source, and peptides were separated with a binary buffer system of buffer A (0.1% formic acid (FA) plus 5% DMSO) and buffer B (80% acetonitrile plus 0.1% FA plus 5% DMSO), at a flow rate of 250 nl/min. The mass spectrometer was operated in positive polarity mode with a capillary temperature of 275°C. The samples were acquired with a DIA method established by Bruderer et al.[34]. Briefly, the method consisted of a MS1 scan (m/z= 300-1,650) with an AGC target of 3x10^6 and a maximum injection time of 60 ms (R= 120,000). DIA scans were acquired at R=

30,000, with an AGC target of 3x10^6, 'auto' for injection time and a default charge state of 4. The spectra were recorded in profile mode and the stepped collision energy was 10% at 25%.

High pH reversed-phase fractionation. HepG2 cells were lysed as described in '*Cell culture and MS sample preparation'*. 150 µg of total protein was digested with a trypsin/Lys-C mix (Promega, V5071) overnight at 37°C with a 1:50 enzyme to protein ratio. The digestion was stopped by adding two volumes of 99% ethylacetate/1% TFA, followed by sonication for 1 min using an ultrasonic probe device (energy output ~40%). The peptides were desalted using 30 mg (8B-S029-TAK) Strata-X-C cartridges (Phenomenex) as follows: a) conditioning with 1 ml of isopropanol; b) conditioning with 1 ml of 80% ACN/5% NH_4OH; c) equilibration with 1 ml of 99% ethylacetate/1% TFA; d) loading of the sample; e) washing with 2 x 1 ml of 99% ethylacetate/1% TFA; f) washing with 1 ml of 0.2% TFA; g) elution with 2 x 1 ml of 80% ACN/5% NH_4OH. The eluates were snap-frozen in liquid nitrogen and lyophilized overnight. The lyophilized peptides were resuspended in 400 µl 0.1 % FA and fractionated using a 3x250 mm xBridge column (Waters) on an ÄKTA HPLC system (GE Healthcare). Fractionation was performed with a flow rate of 0.5 ml/min and with a constant flow of 10% 25 mM ammonium bicarbonate, pH 10. Peptides were separated using a linear gradient of ACN from 7% to 30% over 15 min, followed by a 5-min increase to 55% ACN and a subsequent ramping to 100% ACN. Fractions were collected at 50-sec intervals in 15 ml Falcon tubes to a total of 36 fractions and then pooled to obtain 12 fractions (A1-B1-C1, A2-B2-C2 etc.). All fractions were acidified by addition of FA to a final amount of 0.1% and then lyophilized. Peptides were subsequently resuspended in 100 µl 0.1% TFA and desalted using in-house prepared C18 STAGE tips[43] as follows: a) equilibration with 100 µl isopropanol, b) Equilibration with 100 µl 0.1% TFA, c) loading of the sample, d) washing with 100 µl 0.1% formic acid (FA), e) elution with 30 µl of 80% Acetonitrile/0.1% FA. Peptides were speed-vac dried, resupended in 20 µl 0.1% FA and the concentration estimated on a nanodropTM device (Thermo Fisher Scientific). The samples were then adjusted to 0.4 µg/µl with 0.1% FA, of which 2 µl (800 ng) were injected into the mass spectrometer.

HeLa data with varying gradients

High-pH reversed phase peptide fractionation. 6 µg of HeLa peptides were loaded onto a Waters BEH130 C18 2.1 × 250 mm column in 90 µL of MS loading buffer at a flow rate of 0.5 mL/min using a Dionex Ultimate 3000 HPLC, and column temperature was maintained at

50°C. After loading, a binary gradient of 10% buffer A (2% acetonitrile, 10 mM ammonium formate pH 9) to 40% buffer B (80% acetonitrile, 10 mM ammonium formate pH 9) was formed over 4.4 minutes, followed by a wash-out from 40–100% buffer B over 1 minute, after which the column was held at 100% buffer B for 10 minutes prior to re-equilibration. Fractions were collected over a period of 6.4 minutes from the first peptide elution, with fraction collection each 8 seconds and automatic concatenation into 16 fractions (200 µL fraction volume). Fractions were dried down in a vacuum concentrator (Eppendorf) and resuspended in MS loading buffer (0.3% TFA, 2% acetonitrile).

MS analysis. Peptides were loaded onto a 40 cm column with a 75 µM inner diameter, packed in-house with 1.9 µM C18 ReproSil particles (Dr. Maisch GmbH). Column temperature was maintained at 60°C with a column oven (Sonation GmbH). A Dionex U3000 RSLC nano HPLC system (Thermo Fisher Scientific) was interfaced with a Q Exactive HF X benchtop Orbitrap mass spectrometer (Thermo Fisher Scientific) using a NanoSpray Flex ion source (Thermo Fisher Scientific). For all samples, peptides were separated with a binary buffer system of 0.1% (v/v) formic acid (buffer A) and 80% (v/v) acetonitrile/0.1% (v/v) formic acid (buffer B) and peptides eluted at a flow rate of 400 nl/min. Gradient ranges and durations were as follows: 5–40% buffer B over 30 minutes (DDA library); 3–19% buffer B over 10 minutes and 19–41% over 5 minutes (15 min DIA gradient); 3–19% buffer B over 20 minutes and 19–41% over 10 minutes (30 min DIA gradient); 3–19% buffer B over 40 minutes and 19–41% over 20 minutes (1 h DIA gradient); 3–19% buffer B over 60 minutes and 19–41% over 30 minutes (1.5 h DIA gradient); 3–19% buffer B over 80 minutes and 19–41% over 40 minutes (2 h DIA gradient). For the DDA library, peptides were analysed with one full scan (350-1,400 m/z, R=60,000 at 200 m/z) with a target of 3e6 ions, followed by up to 20 data-dependent MS/MS scans with HCD (target 1e5 ions, maximum IT 28 ms, isolation width 1.4 m/z, NCE 27%, intensity threshold 3.7e5), detected in the Orbitrap (R=15,000 at 200 m/z). Dynamic exclusion was enabled (15 s). For DIA measurements, peptides were analysed with one full scan (350-1,400 m/z, R=120,000 at 200 m/z) at a target of 3e6 ions, followed by 48 data-independent MS/MS scans spanning 350–975 m/z with HCD (target 3e6 ions, maximum IT 22 ms, isolation width 14 m/z, NCE 25%), detected in the Orbitrap (R=15,000 at 200 m/z).

Three species timsTOF Pro benchmark data

Sample preparation. Human cervix carcinoma cell line HeLa was purchased from the German Resource Centre for Biological Material (Braunschweig, Germany). Cells were cultured in Iscove's Modified Dulbecco Medium (PAN Biotech) supplemented with 10% (v/v) fetal calf serum (FCS; Thermo Fisher Scientific), 1% (v/v) glutamine (Carl Roth) and 1% (v/v) sodium pyruvate (Serva) at 37 °C in a 5% CO_2 environment. A pure culture of the *Saccharomyces cerevisiae bayanus*, strain Lalvin EC-1118 was obtained from the Institut Oenologique de Champagne (Epernay, France). Yeast cells were grown in YPD media as described by Fonslow *et al.*[44]. *Escherichia coli* (TOP10) cells were purchased from Thermo Fisher Scientific and grown in LB liquid medium. After harvesting, cells were lysed adding a urea-based lysis buffer (7 M urea, 2 M thiourea, 5 mM DTT, 2% (w/v) CHAPS). Lysis was promoted by sonication at 4°C for 15 min using a Bioruptor (Diagenode, Liège, Belgium). After cell lysis, protein amounts were determined using the Pierce 660 nm Protein Assay (Thermo Fisher Scientific) according to manufacturer's protocol. Tryptic digestion applying a modified filter-aided sample preparation[45] protocol was performed as described in detail before[46]. To generate the two hybrid proteome samples, tryptic peptides were combined in the following ratios as detailed previously[2,46]. Sample A was composed of 65% w/w human, 30% w/w yeast, and 5% w/w *E. coli* proteins. Sample B was composed of 65% w/w human, 15% w/w yeast, and 20% w/w *E. coli* proteins.

LC MS analysis. Samples were analyzed by LC-MS on a trapped ion mobility spectrometry – quadrupole time of flight mass spectrometer (timsTOF Pro, Bruker Daltonics), which was coupled online to a nanoElute nanoflow liquid chromatography system (Bruker Daltonics) via a CaptiveSpray nano-electrospray ion source. Peptides (corresponding to 200 ng) were separated on a reversed-phase C18 column (25 cm x 75 µm i.d., 1.6 µm, IonOpticks, Australia). Mobile phase A was water containing 0.1% (v/v) formic acid, and mobile phase B acetonitrile containing 0.1% (v/v) formic acid. Peptides were separated running a gradient of 2–37% mobile phase B over 100 min at a constant flow rate of 400 nL/min. Column temperature was controlled at 50°C. MS analysis of eluting peptides was performed in diaPASEF mode. For diaPASEF, we adapted the instrument firmware to perform data-independent isolation of multiple precursor windows within a single TIMS separation (100 ms). We used a method with two windows in each 100 ms diaPASEF scan. Sixteen of these scans covered the diagonal scan

line for doubly charged and triply charged peptides in the *m/z* – ion mobility plane with narrow 25 *m/z* precursor windows resulting in a total cycle time of 1.6 s.

BoxCar DIA HEK data

Cell Culture and MS Sample preparation. HEK293 cells were grown in DMEM supplemented with penicillin, streptomycin and 10% FCS. Cells were washed twice with ice-cold PBS, before scraping in PBS and centrifuged at 300 x g for 6 mins at 4°C. Supernatant was aspirated and the pellet lysed in 2.5 % SDS buffered with 50 mM Tris pH 8.1, and heated to 95C for 5 minutes, prior to probe sonication. The BCA assay was used to quantify the protein content of centrifuge-clarified lysates prior to precipitation with 5 volumes of acetone. Pellets were resuspended in 50 mM Tris pH 8.1 containing 8 M urea, reduced with 1 mM DTT and alkylated with 5 mM IAA prior to initiation of digestion overnight with LysC at an enzyme to protein ratio of 1:100. The digest mixture was diluted 4-fold, and trypsin was added at an enzyme to protein ratio of 1:100 for 6 hours, followed by an additional aliquot of trypsin overnight. Digestion was stopped by acidification to 1% TFA, placed on ice for 5 minutes and centrifuged to remove insoluble material. Peptides were desalted with a mixed-mode SPE cartridges (Strata-XC, Phenomenex), activated with 100% methanol, conditioned with 80% Acetonitrile, 0.1% TFA and equilibrated with 0.2% TFA, which was followed by sample loading, washing with 99.9% isopropanol 0.1% TFA, washing twice with 0.2% TFA, and washing once with 0.1% formic acid, before elution with 60% acetonitrile 0.5% ammonium hydroxide. Eluate was flash frozen and dried by centrifugal evaporation.

Offline peptide fractionation. Peptides were resuspended in buffer A (10 mM ammonium bicarbonate) and injected onto a 4.6 x 250 mm 3.5µm Zorbax 300 Extend-C18 column. Peptides were separated on a non-linear gradient exactly as described (Mertins et al., 2018, Nature protocols), using the following composition of buffer B (10 mM ammonium bicarbonate, 90 % acetonitrile). Peptide fractions were frozen at -80 °C before centrifugal evaporation. Peptides were resuspended in 1% TFA, and concatenated at by combining every 24^{th} fraction for the library, or every 8^{th} fraction for the fractionated BoxCar DIA runs, using fractions 13 – 90.

Concatenated or non-fractionated samples were desalted with SEP-PAK tC18 SPE cartridges (Waters), activated with 100 % methanol, conditioned with 80 % acetonitrile, 0.1% TFA, and

equilibrated with 0.2 % TFA. Following sample loading, cartridges were wash with 0.5, 1, and 3 cartridge volumes of 0.2 % TFA, and eluted with 1 volume of 80% acetonitrile, 0.1 % TFA, then frozen before drying in a centrifugal evaporator.

1ug of peptide were loaded onto an Aurora 25cm x 75µm ID, 1.6µm C18 column (Ionopticks) maintained at 40°C. Peptides were separated with an EASY-nLC 1200 system at a flow rate of 300 nl/min using a binary buffer system of 0.1% formic acid (Buffer A) and 80% acetonitrile with 0.1% formic acid (Buffer B), in a two-step gradient from 3-27% B in 105 min and from 27-40 % B in 15 min. All scans were recorded in the Orbitrap of a Fusion Lumos instrument running Tune version 3.3, equipped with a nanoflex ESI source, operated at 1.6 kV, and the RF lens set to 30%. The scan sequence was initiated with MS1 scans from 350-1650 m/z recorded at 120,000 resolution, with an AGC target of 250%, and maximum injection time of 246 ms. The mass range was divided into 24 segments of variable width, with 3 BoxCar scans (multiplexed targeted SIM scan) isolating 8 segments per scan, comprising every third segment. The segments used were identical to those in the MS2 scans, retaining a 1 m/z overlap between boxes in adjacent scans. The normalized AGC target was 200% per segment, with a maximum injection time of 246 ms. BoxCar scans were also recorded at a resolution of 120,000. This was followed by 24 MS2 scans from 200 – 2000 m/z with windows as previously described (Bruderer et al., 2017 MCP). Fragmentation was induced with HCD using stepped collision energy of 22, 27, and 32% for the window center. Each MS2 scan was recorded at a resolution of 30,000, and an AGC target of 1000 % with a maximum injection time of 60 ms.

Data downloads

In addition to the data measured for this publication, we downloaded the following publicly available datasets. The four-species mixture dataset[34] containing *H. sapiens*, *C. elegans*. *S. cerevisiae* and *E. coli* with ratios of 0%, 10%, 20% and 30%, respectively, between replicate groups was downloaded from ProteomeXchange (PXD005573). SCIEX TripleTOF 6600 three species benchmark data[2] was obtained from ProteomeXchange (PXD002952). The HepG2 RNA-seq data is part of the ENCODE dataset[47] and was downloaded from SRA (SRP014320). The HEK RNA-seq data is part of the Cell Atlas dataset[48] and was downloaded from SRA (SRP017465).

Data analysis

In all MaxQuant analyses for generating libraries and for analyzing DIA samples (MaxDIA) version 2.0.0 was used and for all parameters the default values were used unless stated otherwise. Searches were performed with the following FASTA files from UniProt: UP000005640_9606 (*H. sapiens*), UP000007305_4577 (*Z. mays*), UP000002311_559292 (*S. cerevisiae*), UP000000625_83333 (*E. coli*), UP000001940 (*C. elegans*). Methionine oxidation and protein N-terminal acetylation were used as variable modifications in all searches, as is default in MaxQuant.

Comparing number of proteins between datasets. Proteins are assembled into protein groups for identification to account for the redundancy of protein sequences with regards to the peptide evidence distinguishing them. This works in MaxDIA in exactly the same way as in the standard DDA usage of MaxQuant. These protein groups are dataset dependent and hence comparisons between two protein groups tables, for instance in Venn diagrams, or between a protein groups table and RNA-seq data are nontrivial. Here, we follow the route of mapping all protein identifiers in a protein group to Entrez gene identifiers[49]. In the vast majority of cases, protein groups map to single gene identifiers. For cases, in which they map to more than one, both gene identifiers are taken into the set. For counting protein group identifications, we always remove protein groups that are flagged as 'reverse' or 'only identified by site'. For human datasets, we removed protein groups denoted as 'potential contaminant' only if they are of non-human origin and kept human proteins, which consist mostly of human keratins. For the dataset containing bovine plasma the proteins in the standard MaxQuant contaminant list of bovine origin were not removed.

FDR curves. For estimating external FDR, we used a combination of human and maize libraries from reference[34] or of human and maize predicted libraries in discovery mode on the human HepG2 DIA samples. For analyzing library-to-DIA-sample matches and peptide identifications in Fig. 4, we do not apply a protein level FDR and scan through the library-to-DIA-sample FDR. It is crucial to take this approach, in particular when comparing numbers of identifications with other software, since when applying protein-level FDR in MaxQuant, peptides which are not mapping to a protein identified at the specified protein FDR are discarded, unlike in most other software packages. For obtaining the protein-level FDR curves

in Fig.4 we applied a library-to-DIA-sample match FDR of 1%. Peptides that are shared between human and maize proteins were discarded.

RNA-seq data analysis. Raw reads were filtered using trimmomatic[50] (version 0.36) using default parameters for paired-end data. Filtered reads were mapped to the human reference genome GRCh38 (Ensemble release 100) using STAR[51] aligner (version 2.5.3a). Further processing – sorting, converting from SAM to BAM format and indexing – was done using SAMtools[52] (version 1.6). Gene expression quantification (RPKM) for protein-coding genes was performed in Perseus[53] (version 1.6.14.0).

Spectronaut analysis. Raw MS data were processed using Spectronaut version 13.10.191212 using default settings, using a spectral library generated by searching using MaxQuant version 1.6.10.43.

Software development, requirements, availability and usage

MaxDIA has been developed in conjunction with MaxQuant in C#, runs on Windows and Linux operating systems and requires .NET Core 2.1. In addition, .NET Framework 4.7.2 has to be installed on Windows. The graphical user interface version is currently restricted to Windows. A platform-neutral command line version is available. MaxQuant is efficiently running in parallel on arbitrarily many CPUs on single-node platforms. Having 4Gb of memory per parallel running thread is recommended. Disk space should be at least twice the space that is used by the raw data. MaxQuant is freeware and the code is partially open and available at https://github.com/JurgenCox/compbio-base. MaxQuant including MaxDIA can be downloaded from https://www.maxquant.org/. MaxDIA is included in the standard MaxQuant release from version 2.0.0 onward. (MaxQuant 2.0.0 is included in the PRIDE submission for the reviewers.) How to use MaxDIA in library or discovery mode is described in the accompanying Supplementary Notes document. It also contains a list of all user-definable parameters with a description of their meaning.

PRIDE support

We support complete submissions to the PRoteomics IDEntifications (PRIDE) database[28] for the DIA identification results. We extended the mzTab format[29] to cover DIA data sets. For

this purpose, new controlled vocabulary terms were introduced along with additional external reference files. These external reference files contain DIA library matches with mass, intensity and annotation information in a spectral library format (msp-format). MaxQuant will generate a new output folder called 'combined\msp' into which these results are written. A user must provide this folder in addition to raw and mzTab files during submission to PRIDE. More details on a complete PRIDE submission are provided in the Supplementary Notes. This is the first instance of complete PRIDE submissions for DIA data sets.

Data availability

The MS proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifiers PXD022582 (DDA data, login/password for review: reviewer_pxd022582@ebi.ac.uk / oKBHzhLq) and PXD022589 (DIA data, containing also MaxQuant version 2.0.0, login/password for review: reviewer_pxd022589@ebi.ac.uk / yui5MuP8).

ACKNOWLEDGEMENTS

We thank Roland Bruderer for providing data, Georgina D. Barnabas for testing, and all members of the Computational Systems Biochemistry Research Group for helpful discussions. This project was partially funded by the German Ministry for Science and Education (BMBF) funding action MSCoreSys, reference number FKZ 031L0214D and 031L0217A and the Deutsche Forschungsgemeinschaft (SFB1292 Z02, to S.T.). S.Y. is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 792536. Y.P.R. is supported by BBSRC, grant number BB/P024599/1. D.I. is a Chan Zuckerberg Biohub Fellow.

AUTHOR CONTRIBUTIONS

P.S., H.H., F.S.S., N.P., C.W., Ş.Y., J.D.R. and J.C designed and developed the code. D.I., S.T., N.N. S.J.H. and J.C. conceptualized the wet-laboratory experiments and mass spectrometric measurements. D.I., F.M., M.S., U.O., U.D., S.K.S., S.T. and S.J.H. carried out the wet-laboratory experiments and mass spectrometric measurements. H.H., Ş.Y and Y.P.R. designed

and developed the PRIDE support, P.S., H.H., F.S.S. N.P. C.W. S.J.H. and J.C. analyzed the data, M.S., D.I., U.D., N.N., S.J.H. and J.C. wrote online Methods sections, J.C. wrote the manuscript and directed the project.

COMPETING FINANCIAL INTERESTS

The authors state that they have potential conflicts of interest regarding this work: M.S. and U.O. are employees of Evotec, N.N. and S.K.S. are employees of Bruker and J.D.R. is employee of Bosch.

REFERENCES

1. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35–35 (2014).
2. Navarro, P. *et al.* A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3685
3. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
4. Azvolinsky, A., DeFrancesco, L., Waltz, E. & Webb, S. 20 years of Nature Biotechnology research tools. *Nat. Biotechnol.* **34**, 256–261 (2016).
5. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **1**, 207–234 (2018).
6. Sinitcyn, P. *et al.* MaxQuant goes Linux. *Nat. Methods* **15**, 401 (2018).
7. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* (2014). doi:10.1038/nbt.2841
8. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
9. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* (2015). doi:10.1074/mcp.M114.044305
10. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* (2020). doi:10.1038/s41592-019-0638-x
11. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513–2526 (2014).
12. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat Meth* **14**, 921–927 (2017).
13. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
14. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
15. Tiwary, S. *et al.* High quality MS/MS spectrum prediction for data-dependent and -independent acquisition data analysis. *Nat Methods* (2019).

16. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* (2019). doi:10.1038/s41592-019-0426-7
17. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* (2020). doi:10.1038/s41467-019-13866-z
18. Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* (2020). doi:10.1038/s41467-020-15346-1
19. Lou, R. *et al.* Hybrid Spectral Library Combining DIA-MS Data and a Targeted Virtual Library Substantially Deepens the Proteome Coverage. *iScience* (2020). doi:10.1016/j.isci.2020.100903
20. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* (2019). doi:10.1038/s41592-018-0260-3
21. Graves, A. *et al.* A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 855–868 (2009).
22. Chen, T. & Guestrin, C. XGBoost : Reliable Large-scale Tree Boosting System. *arXiv* (2016). doi:10.1145/2939672.2939785
23. Prianichnikov, N. *et al.* MaxQuant software for ion mobility enhanced shotgun proteomics. *bioRxiv* (2019). doi:10.1101/651760
24. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* (2018). doi:10.1038/s41592-018-0003-5
25. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobil. Spectrom.* (2011). doi:10.1007/s12127-011-0067-8
26. Silveira, J. A., Ridgeway, M. E. & Park, M. A. High resolution trapped ion mobility spectrometery of peptides. *Anal. Chem.* (2014). doi:10.1021/ac501261h
27. Meier, F. *et al.* Online parallel accumulation – serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer.
28. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1106
29. Griss, J. *et al.* The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Mol. Cell. Proteomics* **13**, 2765–2775 (2014).
30. Martens, L. *et al.* mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **10**, R110 000133 (2011).
31. Cox, J., Michalski, A. & Mann, M. Software lock mass by two-dimensional minimization of peptide mass errors. *J Am Soc Mass Spectrom* **22**, 1373–1380 (2011).
32. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
33. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
34. Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. Cell. Proteomics* mcp.RA117.000314 (2017). doi:10.1074/mcp.RA117.000314
35. Ludwig, C. *et al.* Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial . *Mol. Syst. Biol.* (2018). doi:10.15252/msb.20178126
36. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628 (2008).
37. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs.

Nature **455**, 58–63 (2008).
38. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* (2020). doi:10.1038/s41576-020-0258-4
39. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
40. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* (2011). doi:10.1038/msb.2011.28
41. Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.* (2020). doi:10.1038/s41467-020-14609-1
42. Müller, F., Kolbowski, L., Bernhardt, O. M., Reiter, L. & Rappsilber, J. Data-independent acquisition improves quantitative cross-linking mass spectrometry. *Mol. Cell. Proteomics* (2019). doi:10.1074/mcp.TIR118.001276
43. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**, 663–670 (2003).
44. Fonslow, B. R. *et al.* Digestion and depletion of abundant proteins improves proteomic coverage. *Nat. Methods* (2013). doi:10.1038/nmeth.2250
45. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* (2009). doi:10.1038/nmeth.1322
46. Distler, U., Kuharev, J., Navarro, P. & Tenzer, S. Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics. *Nat. Protoc.* (2016). doi:10.1038/nprot.2016.042
47. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
48. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-.).* (2017). doi:10.1126/science.aal3321
49. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkq1237
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu170
51. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp352
53. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–40 (2016).

FIGURE LEGENDS

Fig. 1: Overview of the MaxDIA workflow. MaxDIA can be operated in library and discovery mode. Many concepts and algorithms, for instance for protein quantification, are re-used from the conventional MaxQuant workflow for DDA data and have been further developed for DIA. This results in an end-to-end DIA software that contains many established MaxQuant concepts, like label-free quantification with MaxLFQ or iBAQ quantification.

Fig. 2: 3D/4D feature detection of precursors and fragments. a, Visualization of precursor and fragments of a peptide measured on an Orbitrap. The raw data can be visualized together with the peak detection results as heat maps and 3D models for precursor and fragment data in the graphical user interface of MaxQuant. **b,** Two peptides with nearly equal mass, both with charge 2 and having very similar retention times are resolved by ion mobility on a timsTOF Pro mass spectrometer. A heat map visualizes intensities as a function of retention time and collision cross section for the precursor isotope patterns. The two respective MS/MS spectra of fragments assigned to the precursors are shown.

Fig. 3: Performance evaluation. 27 technical replicates of HepG2 cell lysate were analyzed on an Orbitrap mass spectrometer (see online Methods). **a,** Number of identified protein groups with 1% FDR on protein and peptide level, and number of peptides at 1% library-to-DIA-sample FDR obtained with MaxDIA and Spectronaut. **b,** Histograms of peptide lengths identified with MaxDIA (blue) and in Spectronaut (red). **c,** Number of proteins with at most x out of 27 valid values for Spectronaut (red), MaxDIA with MaxLFQ minimum ratio count = 1 (blue, dashed) and = 2 (blue, solid). Multiple curves for the two MaxQuant seies of curves correspond to seven different choices for the transfer q-value (0.01, 0.02, 0.05, 0.1, 0.2 and 0.5). **d,** Histograms of coefficients of variation for analyses with default settings in MaxDIA (blue) and in Spectronaut (red). **e,** Log-log scatter plot of LFQ intensities between two representative replicates obtained with MaxQuant. The two replicates were chosen to have the median Pearson correlation of all pair-wise replicate comparisons. **f,** Same as in panel f for Spectronaut intensities. Similarly, the two replicates were chosen to represent the median Pearson correlation coefficient of all pair-wise comparisons. **g,** Heat map with all pair-wise Pearson correlations between the 27 replicates for MaxDIA (upper triangle) and Spectronaut (lower traingle). The two values corresponding to the comparisons in panels e and f are marked with red squares. **h,** Log-log scatterplot of iBAQ protein intensities from MaxDIA against

Spectronaut protein intsnsities. **i,** Log-log scatterplot of MaxDIA iBAQ values averaged over the replicates against RPKM values from RNA-seq data. **j,** Same as panel i with protein intensities from Spectronaut.

Fig. 4: Internal and external FDR. a, Number of identifications (blue: matches, green: peptides, red: protein groups) as a function of estimated FDR. The FDR is once estimated with the 'internal' target-decoy method implemented in MaxQuant (solid lines) and once with the 'external' method using mixing maize and human samples for generating the library and using only human sample in the DIA runs (dashed lines). **b,** Same as in panel a but using *in-silico* predicted libraries generated using DeepMass:Prism[15] **c,** Same as panel a but using the raw score instead of the machine learning-derived score. **d,** Same as panel b but using the raw score instead of the machine learning-derived score.

Fig. 5: MaxLFQ for DIA. a, Stacked inter-quartile rages of protein ratio distributions in the small-ratio four-species dataset from Bruderer et al.[34] using different versions of MaxLFQ for DIA and compared to the results from this publication. **b,** Quantification of a three-species benchmark mixture measured on a SCIEX TripleTOF 6600 instrument mixing proteomes from three species in defined ratio[2] with MaxLFQ for DIA. The accompanying DDA library was used. **c,** Same as b, but analyzed with MaxDIA in discovery mode. **d,** Quantification of a three-species benchmark mixture measured on a Bruker timsTOF Pro instrument mixing proteomes from three species in defined ratio using a DDA library. **e,** Same as d, but analyzed in discovery mode.

Fig. 6: BoxCar and fractionated DIA. a, Schedule of libraries and DIA samples. Three different library approaches, single-shot, deep fractionated and discovery mode library were compared to single-shot deep fractionated DIA samples. **b,** MaxLFQ quantification between three replicates of fractionated BoxCar DIA samples analyzed in discovery DIA mode. All pair-wise Pearson correlations are above 0.99. **c,** Venn diagram-like comparison represented as bar plot between RNA-seq data of HEK cells and three different library methods applied to the fractionated DIA samples. All data has been mapped to gene identifiers **d,** Histogram of protein identifications mapped to gene identifiers sorted into bins according to log2 RPKM values of the RNA-seq data.

Figure 1

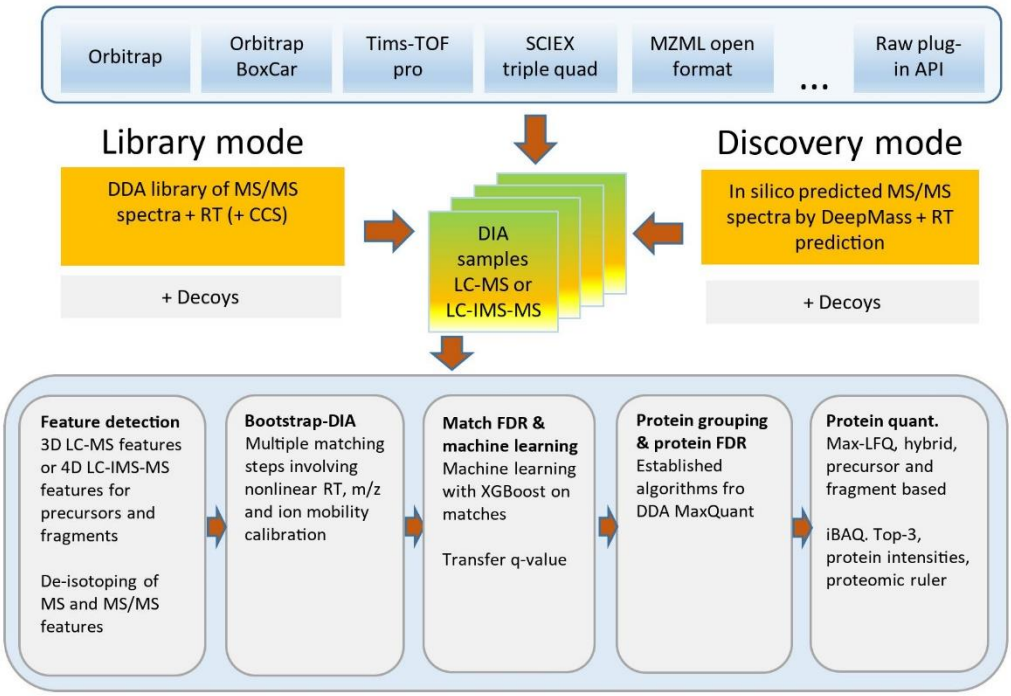

Orbitrap
Orbitrap BoxCar
Tims-TOF pro
SCIEX triple quad
MZML open format
...
Raw plug-in API
Library mode
DDA library of MS/MS spectra + RT (+ CCS)
+ Decoys
DIA samples LC-MS or LC-IMS-MS
Discovery mode
In silico predicted MS/MS spectra by DeepMass + RT prediction
+ Decoys
Feature detection
3D LC-MS features or 4D LC-IMS-MS features for precursors and fragments
De-isotoping of MS and MS/MS features
Bootstrap-DIA
Multiple matching steps involving nonlinear RT, m/z and ion mobility calibration
Match FDR & machine learning
Machine learning with XGBoost on matches
Transfer q-value
Protein grouping & protein FDR
Established algorithms fro DDA MaxQuant
Protein quant.
Max-LFQ, hybrid, precursor and fragment based
iBAQ. Top-3, protein intensities, proteomic ruler

Figure 2

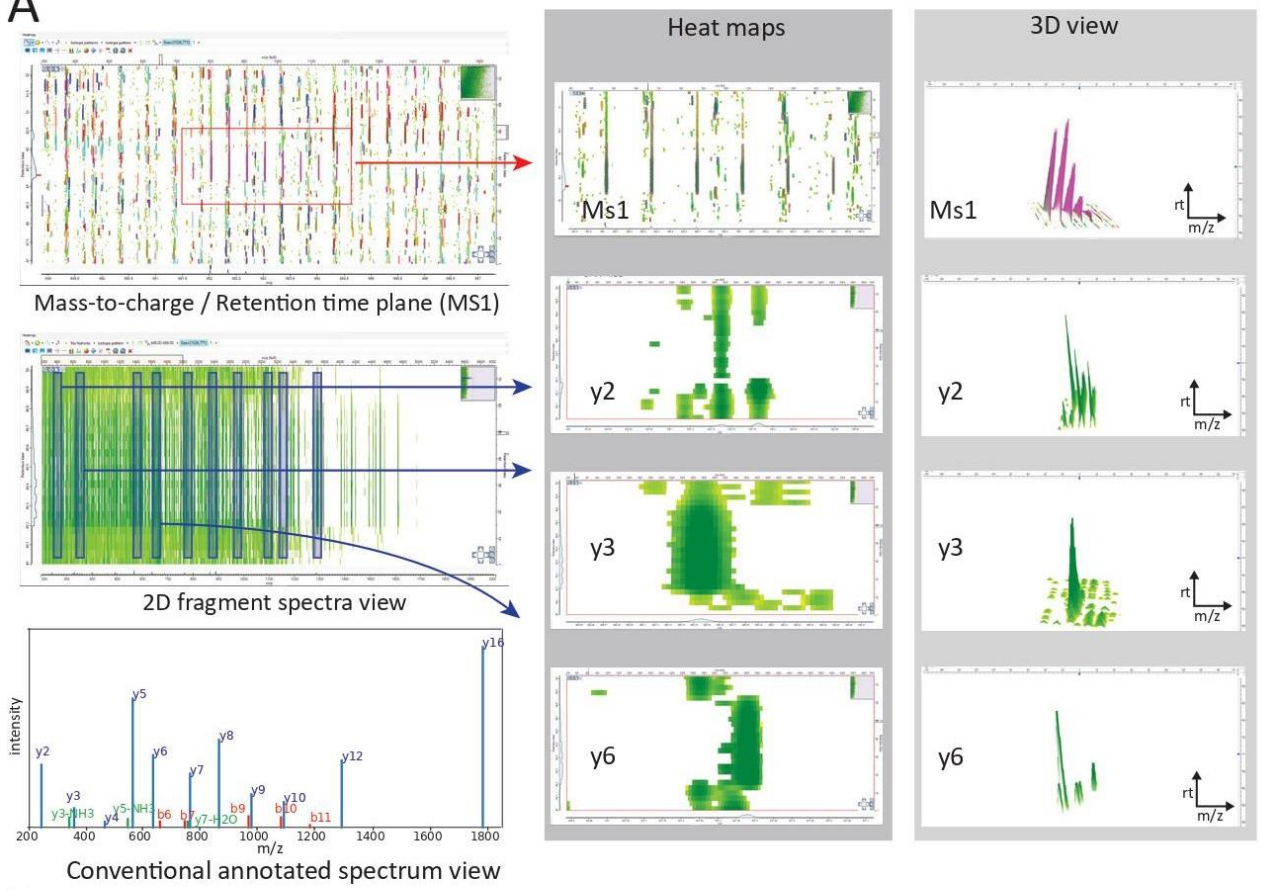

A
Heat maps
3D view
Ms1
Ms1
Mass-to-charge / Retention time plane (MS1)
y2
y2
2D fragment spectra view
y3
y3
y6
y6
intensity
m/z
Conventional annotated spectrum view
B
Mass-to-charge / Ion mobility plane
Conventional spectrum view

Figure 3

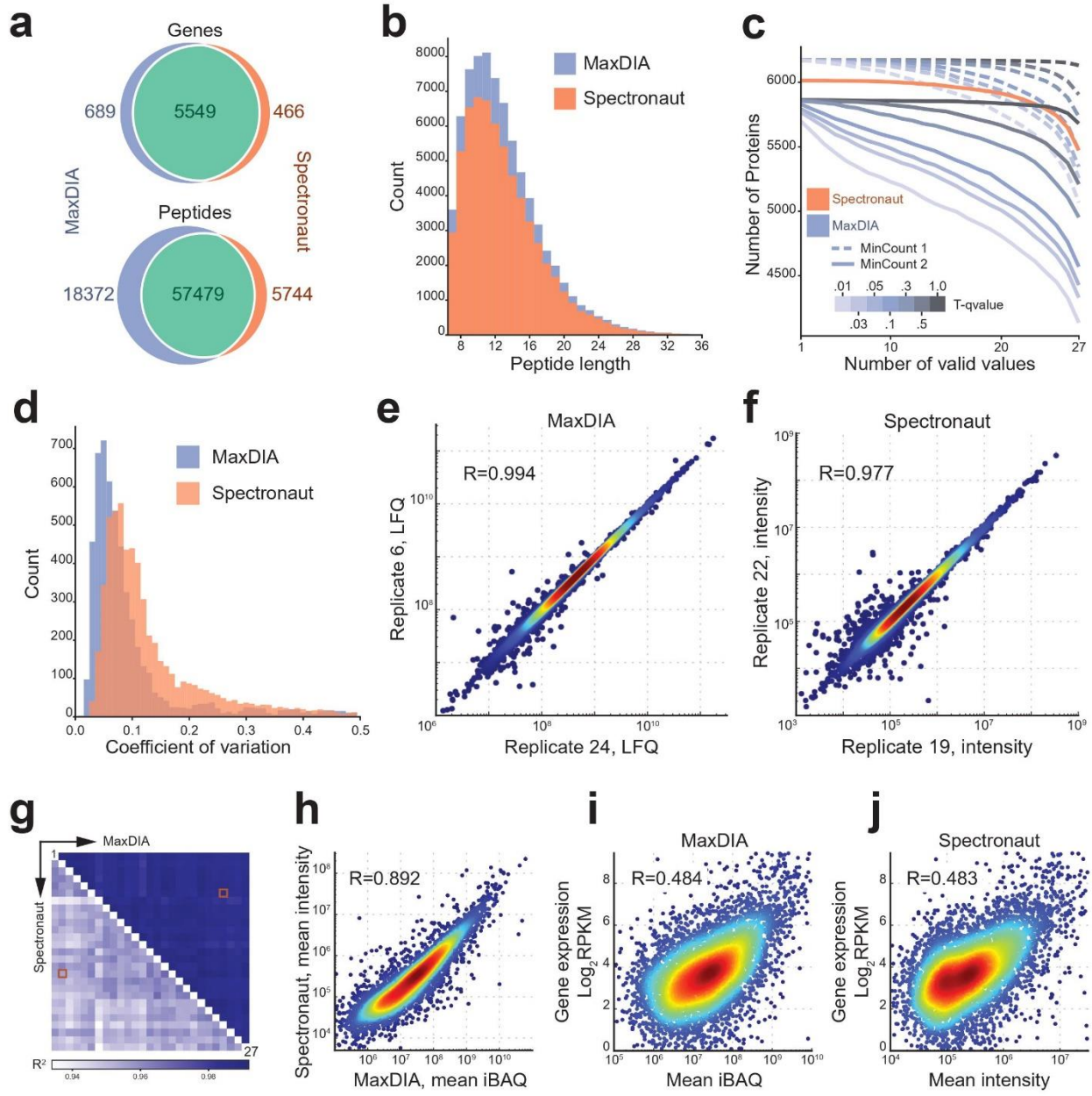

a
Genes
689
5549
466
MaxDIA
Spectronaut
Peptides
18372
57479
5744
b
MaxDIA
Spectronaut
Count
Peptide length
c
Number of Proteins
Spectronaut
MaxDIA
MinCount 1
MinCount 2
T-qvalue
Number of valid values
d
MaxDIA
Spectronaut
Count
Coefficient of variation
e
MaxDIA
R=0.994
Replicate 6, LFQ
Replicate 24, LFQ
f
Spectronaut
R=0.977
Replicate 22, intensity
Replicate 19, intensity
g
MaxDIA
Spectronaut
R²
h
R=0.892
Spectronaut, mean intensity
MaxDIA, mean iBAQ
i
MaxDIA
R=0.484
Gene expression Log2 RPKM
Mean iBAQ
j
Spectronaut
R=0.483
Gene expression Log2 RPKM
Mean intensity

Figure 4

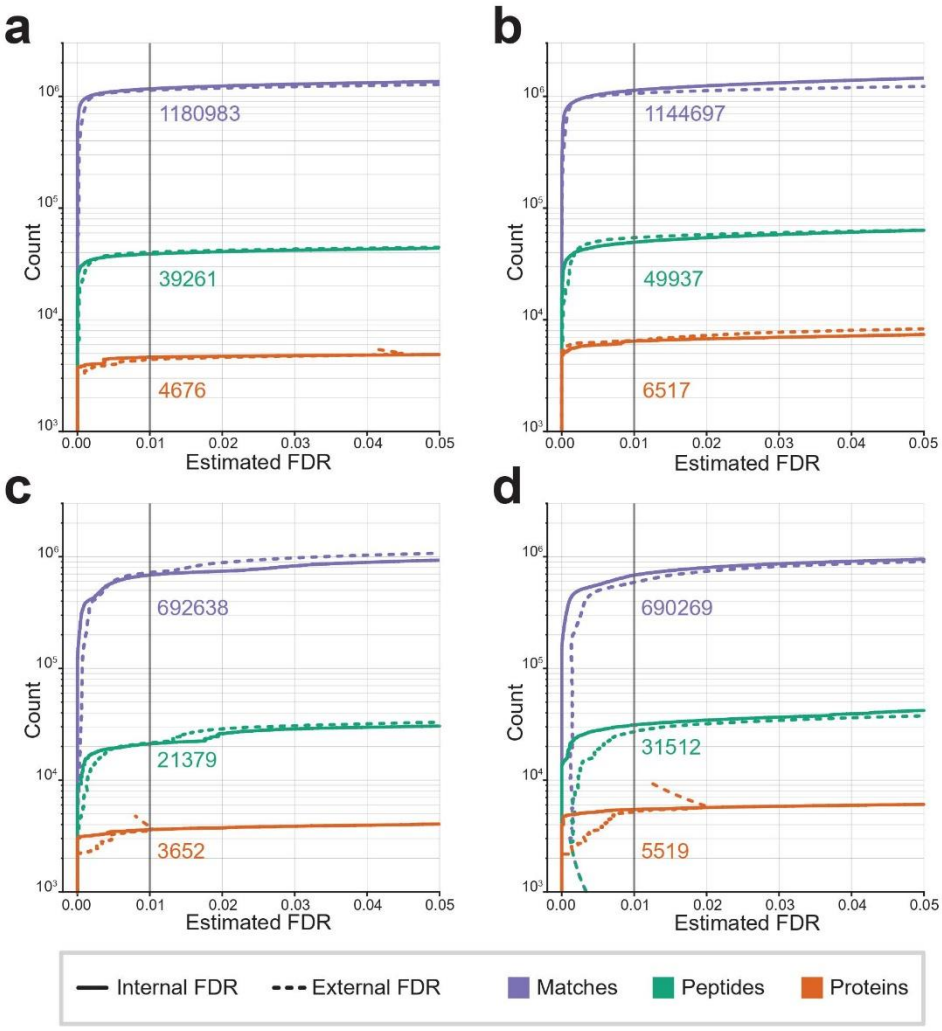

a
b
c
d
Count
Estimated FDR
1180983
39261
4676
1144697
49937
6517
692638
21379
3652
690269
31512
5519
Internal FDR
External FDR
Matches
Peptides
Proteins

Figure 5

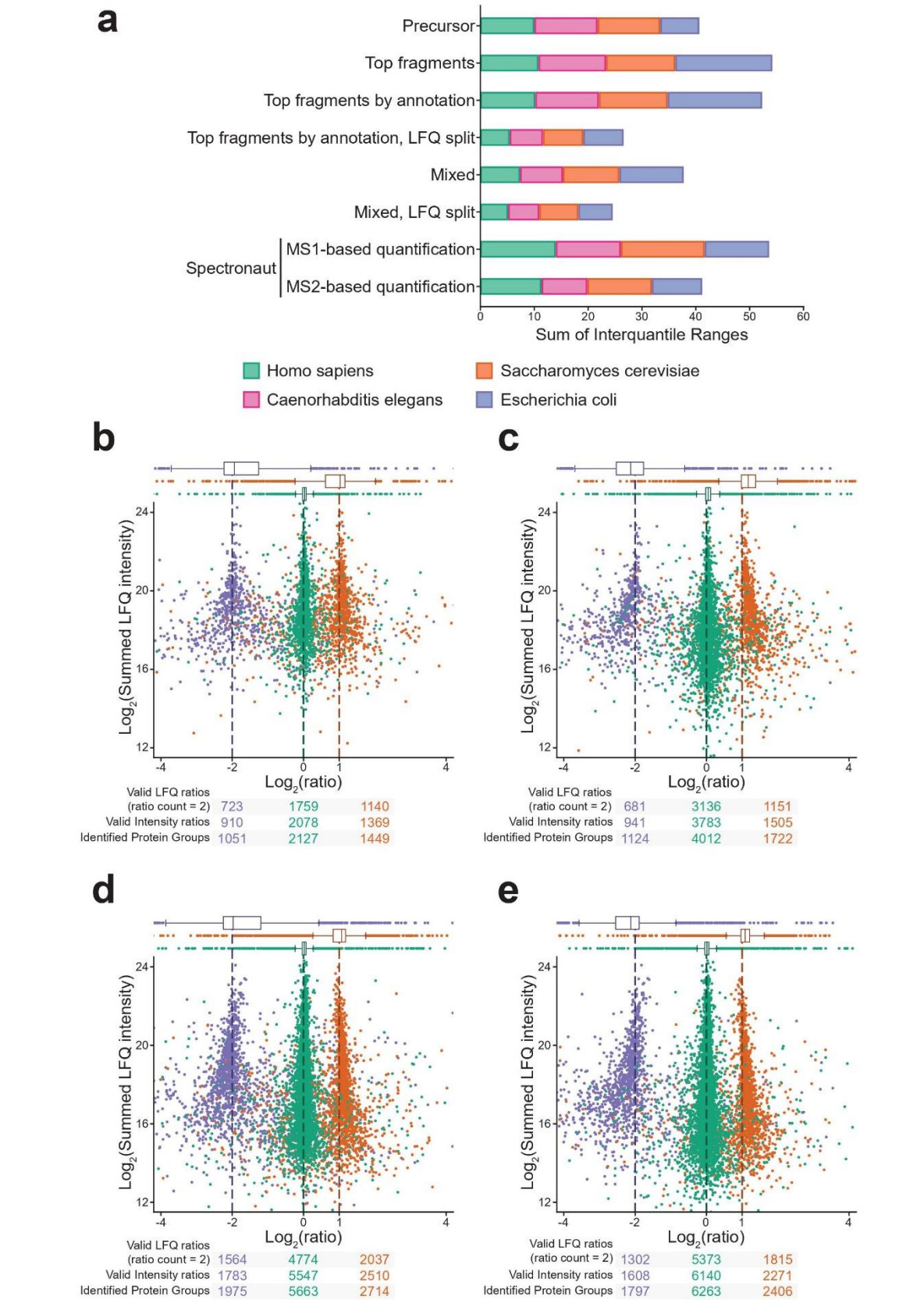
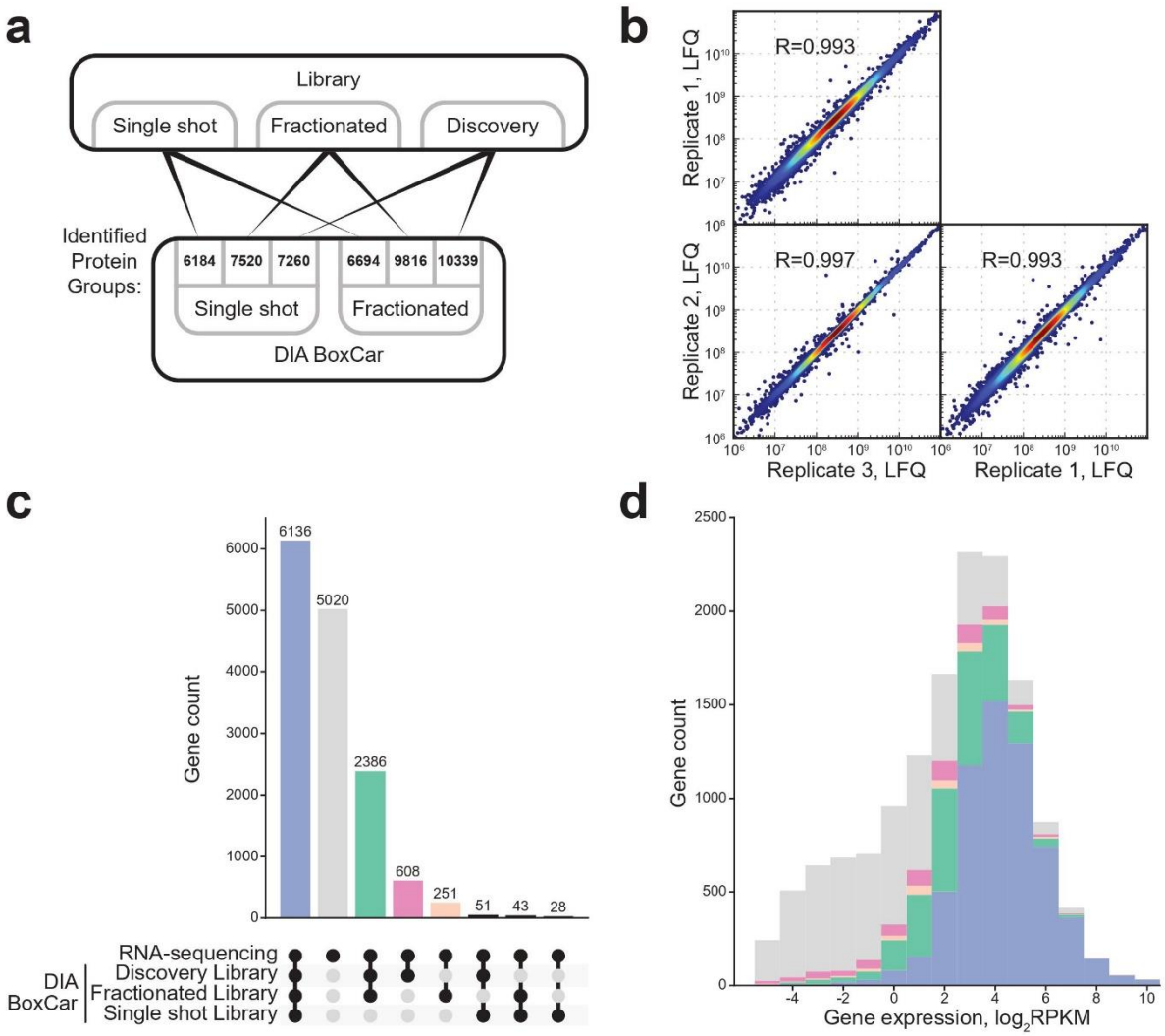
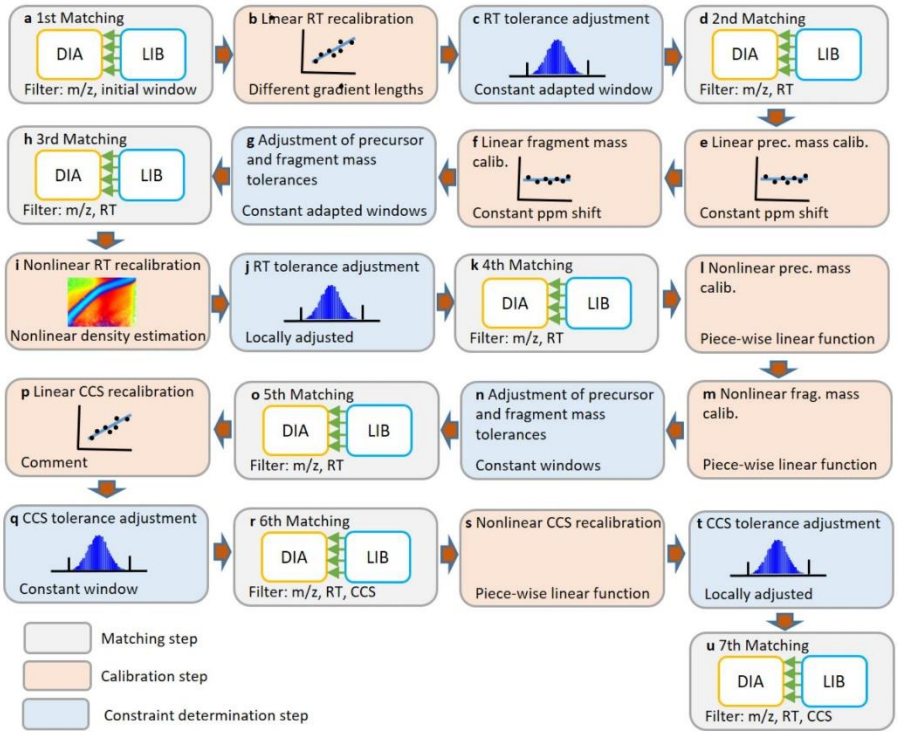

a
Precursor
Top fragments
Top fragments by annotation
Top fragments by annotation, LFQ split
Mixed
Mixed, LFQ split
Spectronaut
MS1-based quantification
MS2-based quantification
0 10 20 30 40 50 60
Sum of Interquantile Ranges
Homo sapiens
Caenorhabditis elegans
Saccharomyces cerevisiae
Escherichia coli
b
c
d
e
Log2(Summed LFQ intensity)
Log2(ratio)
Valid LFQ ratios (ratio count = 2) 723 1759 1140
Valid Intensity ratios 910 2078 1369
Identified Protein Groups 1051 2127 1449
Valid LFQ ratios (ratio count = 2) 681 3136 1151
Valid Intensity ratios 941 3783 1505
Identified Protein Groups 1124 4012 1722
Valid LFQ ratios (ratio count = 2) 1564 4774 2037
Valid Intensity ratios 1783 5547 2510
Identified Protein Groups 1975 5663 2714
Valid LFQ ratios (ratio count = 2) 1302 5373 1815
Valid Intensity ratios 1608 6140 2271
Identified Protein Groups 1797 6263 2406

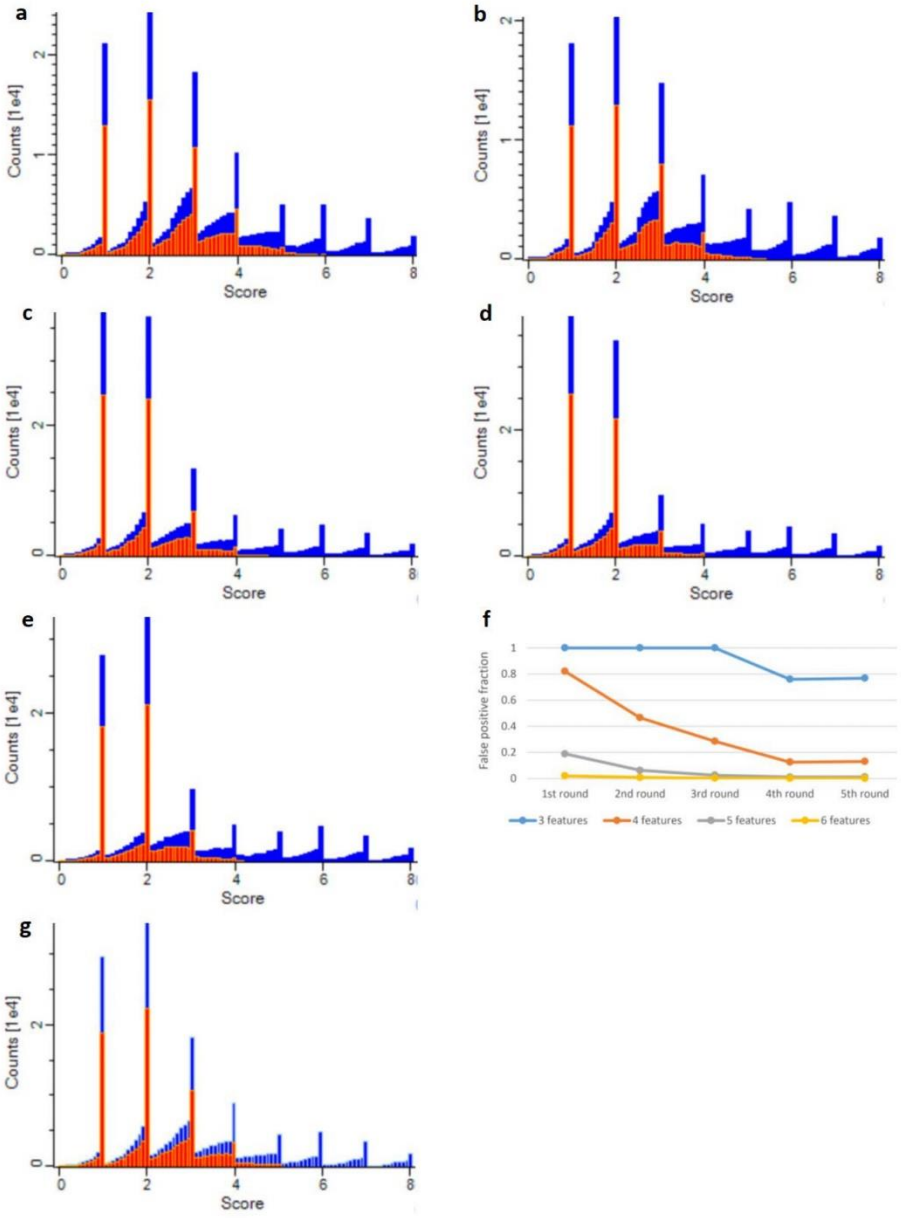
Figure 6


a
Library
Single shot
Fractionated
Discovery
Identified Protein Groups:
6184
7520
7260
6694
9816
10339
Single shot
Fractionated
DIA BoxCar
b
R=0.993
R=0.997
R=0.993
Replicate 1, LFQ
Replicate 2, LFQ
Replicate 3, LFQ
Replicate 1, LFQ
c
Gene count
6136
5020
2386
608
251
51
43
28
RNA-sequencing
DIA BoxCar
Discovery Library
Fractionated Library
Single shot Library
d
Gene count
Gene expression, log2RPKM



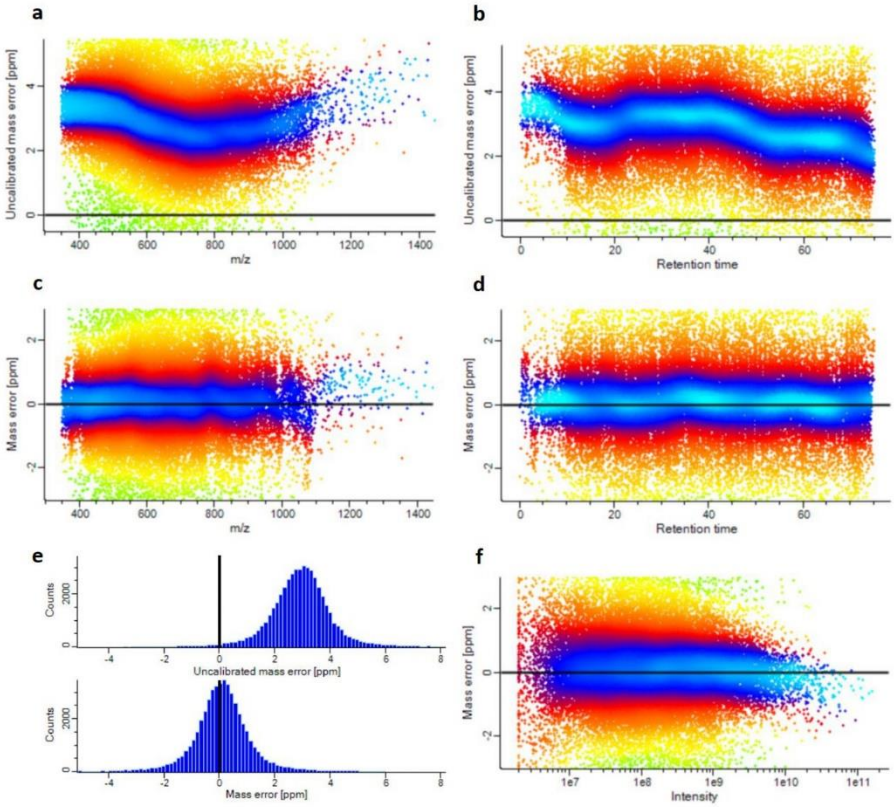
Supplementary Fig. 1: The bootstrap DIA workflow. This sequence of algorithmic steps is applied to each DIA sample vs. the whole library. A matching step is usually followed by a step in which a calibration function (e.g. precursor m/z recalibration function) is determined from the matches found in the previous step. Then constraints (e.g. m/z deviation windows) are updated for the next round of matching. The DDA samples constituting the library are assumed to be retention time (and ion mobility if applicable) aligned to each other. **a,** The first matching from the library spectra to the DIA sample is performed with initial m/z windows for precursor and fragments of 20 p.p.m. by default and without restrictions on retention times or collision cross sections. **b,** Based on these matches, a linear recalibration is calculated to adjust for different total gradient lengths of library and DIA samples. **c,** After the linear retention time calibration has been calculated and applied, a time window is calculated from the data, which defines the allowed retention time difference for the next step. **d,** The second matching still uses the initial m/z windows and in addition uses the time window determined in the previous step. **e,** Based on the matches of the previous step a linear precursor m/z shift in p.p.m. between the DIA sample and calculated peptide masses is determined. **f,** Similarly, a fragment m/z shift is calculated from the data. **g,** Next, precursor and fragment m/z tolerances are calculated based on the

distributions of m/z differences between DIA sample and theoretically calculated masses. **h,** The third matching uses adapted m/z and retention time windows which are applied to the linear calibrated data. **i,** The elimination of noise achieved by the adapted tolerances used in the matching in the previous step allows now to perform nonlinear retention tine calibration. **j,** A time dependent nonlinear allowed region is determined from the data. **k,** The fourth matching uses more stringent retention time constraints than the third matching, since it is applied to nonlinear calibrated data. **l,** Now a nonlinear function for the calibration of precursors is determined from the data. **m.** Similarly, fragment m/z are nonlinear recalibrated. **n,** New, more stringent precursor and fragment m/z tolerances are calculated from the distributions of mass errors. **o.** Another matching step with updated constraints is performed. **p,** A linear function for the recalibration of CCS values is calculated from the data, in case of ion mobility spectrometry. **q,** A tolerance window for the acceptance of CCS value deviations is calculated. **r,** A matching round with constraints on the CCS values is performed. **s,** A nonlinear CCS calibration function is determined. **t,** CCS tolerance is adapted to the nonlinear calibrated data. **u,** The final round of matching is performed without constraints on retention time and CCS values. Instead, these deviations are used as features in the XGBoost-based machine learning. Precursor and fragment masses are still filtered with hard windows for the deviations.



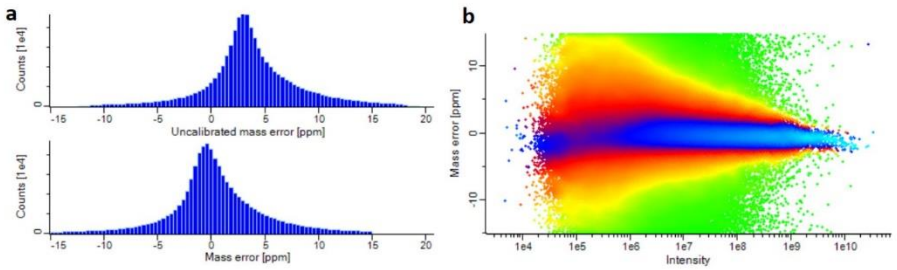
Supplementary Fig. 2: Score distributions along the bootstrap DIA workflow. Histograms of score distributions, separately for target and decoy hits after the different matching steps in the bootstrap DIA workflow. Target (blue) and decoy (red) distributions

are stacked on top of each other. A single run of the HepG2 Orbitrap dataset (DIA_13.raw) was used. **a,** Score histogram after the first matching step. (Step a in Supplementary Fig. 1.) No constraints on the retention time are used. Initial tolerances of 20 p,p,m. are applied to precursor and fragment mass matches. The spikes at integer score values correspond to matches in which all matching fragments hit exactly the apex of the peak in retention time direction. The peaks from one to four matching fragments are dominated by false positives, since these bins have half or even more decoy hits. Score values of six or above indicate correctness of the match since hits are strongly suppressed. **b,** Score histogram after the second matching step. (Step d in Supplementary Fig. 1.) Retention time is filtered after linear retention time calibration between library and DIA sample and after determining a tolerance from the distribution of retention time differences. **c,** Score histogram after the third matching step. (Step h in Supplementary Fig. 1.) Linear ppm shifts are applied to precursor and fragment masses and mass tolerances are adapted accordingly. Scores larger than four indicate few false positives, **d,** Score histogram after the fourth matching step. (Step k in Supplementary Fig. 1.) **e,** Score histogram after the fifth matching step. (Step o in Supplementary Fig. 1.) in which nonlinear mass recalibrations have been applied to the data. **f,** Each profile shows the rate of false positive matches after each of the five different matching steps. **g,** After all recalibrations have been applied, the final matching is done without constraints on retention times, but the mass constraints are kept. (The corresponding score distribution is displayed.) Instead the deviation from the calibrated retention time is offered as a feature to the machine learning for calculating an enhanced score. This strategy (hard mass cutoffs and soft, machine learning based, retention time cutoff) resulted in the highest number of identifications. Similarly, a soft cutoff is used for collision cross sections in ion mobility spectrometry data.

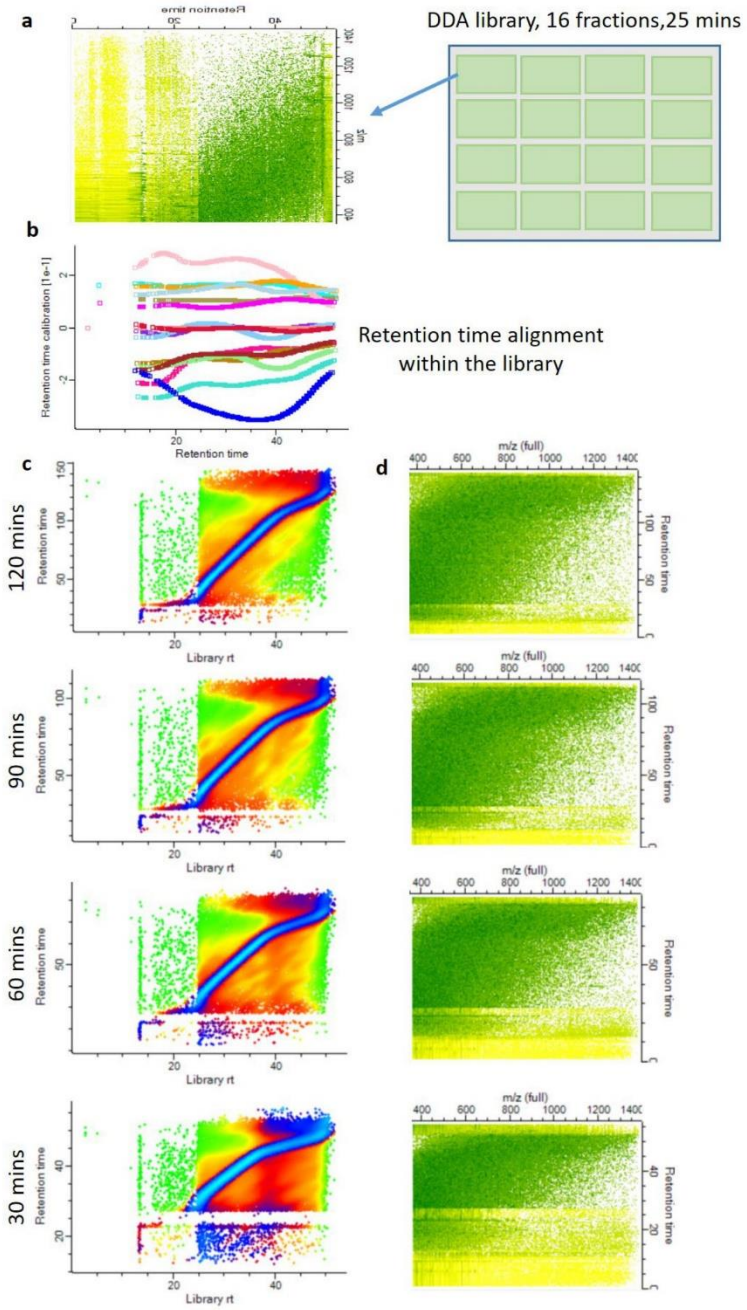


Supplementary Fig. 3: Nonlinear m/z recalibration of precursors. One consequence of the bootstrap DIA is that masses of precursors and fragments are nonlinearly recalibrated against theoretically calculated molecule masses. This replaces the software lock mass strategy used in DDA MaxQuant which is based on a 'first search' with the Andromeda search engine to produce the recalibration curves. We use the same data as in Supplementary Fig. 2 to compare mass errors before and after recalibration. In all panels, data points are color coded according to the conditional data density. For this, the bivariate density of data points is divided by the marginal distribution on the x-axis. Blue signifies the region of highest conditional density. **a,** Mass error in p.p.m. of precursor ions as a function of m/z. **b,** Same precursor mass error as in panel a as a function of retention time. **c,d** Mass errors of panels a and b after recalibration through bootstrap DIA. The high-density regions are centered around 0 error. **e,** Histograms of precursor mass errors before and after recalibration. The medians of the error distributions are at 2.96 p.p.m. before and at 0.099 ppm after recalibration. The FWHM reduces from 1.92 to 1.61 p.p.m.. **f,** Dependency of the precursor mass error on logarithmic intensity. Interestingly, does the

distribution of mass error not depend much on the intensity, since the lines of constant density (constant color) run approximately horizontally.

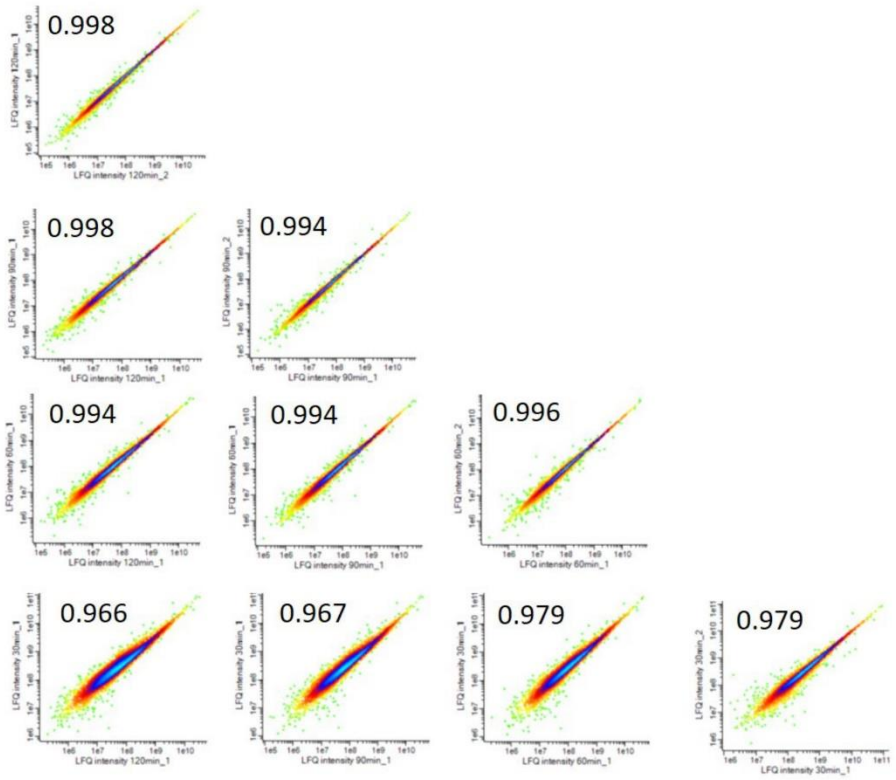


Supplementary Fig. 4: Nonlinear m/z recalibration of fragments. a, Histograms of fragment mass errors before and after recalibration. Since in this dataset, the statistical fluctuations are much larger for the fragment mass errors compared to the precursors, the correction of systematic errors is of less importance here. **b,** Dependency of the fragment mass error on logarithmic intensity. The distribution of mass errors gets wider towards lower intensities.

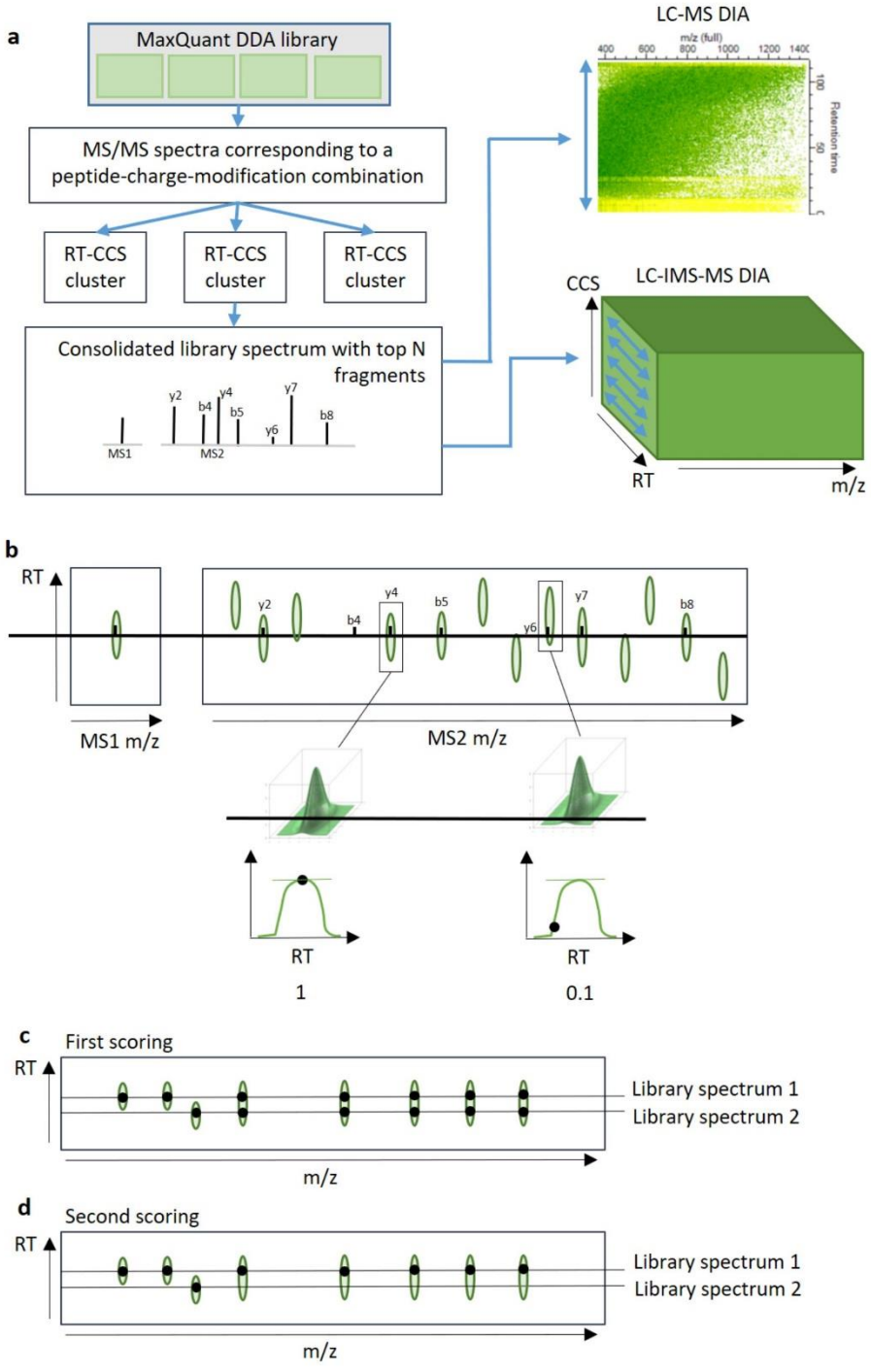

a
DDA library, 16 fractions,25 mins
b
Retention time alignment within the library
Retention time calibration [1e-1]
Retention time
c
d
120 mins
90 mins
60 mins
30 mins
Retention time
Library rt
m/z (full)

7

Supplementary Fig. 5: Nonlinear retention time alignment between different gradients. a, A library of HeLa cell lysate was measured in 16 high-pH reversed phase peptide fractions with an active gradient time of 25 minutes. **b,** While analyzing the library in MaxQuant in DDA mode, retention times are aligned between the LC-MS runs in the library. **c,** Alignment of library retention times against for DIA samples with active gradient times of 120, 90, 60 and 30 minutes. **d,** Heat map views of the MS1 m/z-retention time planes of the respective DIA samples.



Supplementary Fig. 6: Nonlinear retention time alignment: LFQ after the alignment. Triangular matrix of scatter plots showing MaxLFQ quantification results between the four DIA samples with different gradients. The alignment enables precise quantification even between samples with vastly different gradients. On the diagonal, technical replicates with same gradients are shown. Pearson correlation coefficients between logarithmic LFQ intensities range from 0.998 for 120h gradients to 0.979 for 30h gradients. Throughout, quantification between non-equal gradients results in Pearson correlation values close to the one achieved with equal gradients of the respective shorter length.


a
MaxQuant DDA library
MS/MS spectra corresponding to a peptide-charge-modification combination
RT-CCS cluster
RT-CCS cluster
RT-CCS cluster
Consolidated library spectrum with top N fragments
y2
b4
y4
b5
y7
y6
b8
MS1
MS2
LC-MS DIA
m/z (full)
Retention time
LC-IMS-MS DIA
CCS
RT
m/z
b
RT
MS1 m/z
MS2 m/z
RT
1
RT
0.1
c
First scoring
RT
Library spectrum 1
Library spectrum 2
m/z
d
Second scoring
RT
Library spectrum 1
Library spectrum 2
m/z

Supplementary Fig. 7: Scoring library spectra against DIA samples. a, Libraries are collections of DDA samples analyzed with MaxQuant. MS/MS spectra from the library are first sub-divided into unique peptide-charge-modification combinations. Each such combination that has assigned more than one MS/MS spectrum to it is then clustered into retention time clusters. Prerequisite for this is that all library samples are retention-time aligned to each other. The idea is that if a peptide is eluting at more than one place in a gradient, it will be stored as multiple instances in the library with different retention times. This is feasible, since from the MaxQunat DDA analysis it is known how the peptides elute from their MS1 features. For data with ion mobility spectrometry this kind of library feature clustering is done in the two-dimensional space consisting of retention times and collision cross sections. A resulting cluster may still contain more than one MS/MS spectrum. In that case, the one with the highest Andromeda score is chosen. This spectrum is then filtered to the top-N most intense fragment peaks. These are then scored against the DIA sample. By default, is N = 7. We visit each retention time in a DIA LC-MS run and calculate the score which is defined below. The matching position is defined as the retention time at which the highest score is achieved. This highest value of the score is also defined as the matching score of this library spectrum to the DIA sample. For ion mobility spectrometry, this score maximization takes place in the two-dimensional space of all retention time and ion mobility value pairs. **b,** For calculating the score of a library spectrum at a certain retention time (and CCS value) in the DIA sample, one first searches with a given mass tolerance for 3D/4D features that match the precursor and the N (typically = 7) top fragment peaks. For each spectrum mass that matches a feature in the DIA sample we calculate the apex fraction which is the ratio of the maximum peak intensity to the intensity at the current retention time. To obtain the score, we sum up the apex fractions for the precursor (in case one was matched) and the matching fragments. **c,** So far the scoring was done independently for each consolidated library spectrum. This can lead to multiple usages of a DIA feature in several library matches. **d,** To prohibit over-interpretation, we perform a second round of scoring. This time we put the library spectra in descending order according to the score they achieved in the first round of scoring. The same procedure is repeated, but now it is remembered which features in the DIA sample (precursors and fragments) have already been assigned and these will be prohibited from being assigned a second time. Note that a precursor match is not required but contributes the same way to the total score as each fragment does.

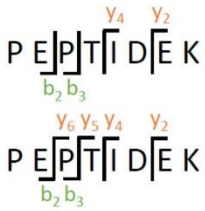
a

No.	Feature
1	Score
2	Pearson intensity correlation
3	Spearman intensity correlation
4	DIA range index
5	Sequence length
6	Charge
7	Retention time
8	Retention time error
9	1/K0
10	1/K0 error
11	Number of (unfiltered) library fragments
12	Andromeda score of library spectrum
13	Number of missed cleavages
14	Number of modifications
15	Sample fragment overlap
16	Number of fragments found
17	Library index (in case of multiple libraries)
18	Precursor mass (calculated)
19	Precursor match score
20	Precursor mass error [ppm]
21	Precursor RT peak length (number of scans)
22	Precursor found (yes/no)
23	Precursor has isotope pattern (yes/no)
24	Mass difference to range min
25	Mass difference to range max

b

No.	Feature
1	Fragment match score
2	Fragment mass error [ppm]
3	Fragment RT peak length (number of scans)
4	Fragment found (yes/no)
5	Fragment has isotope pattern (yes/no)

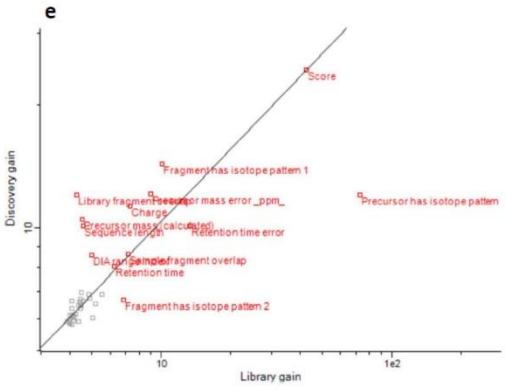
c



d

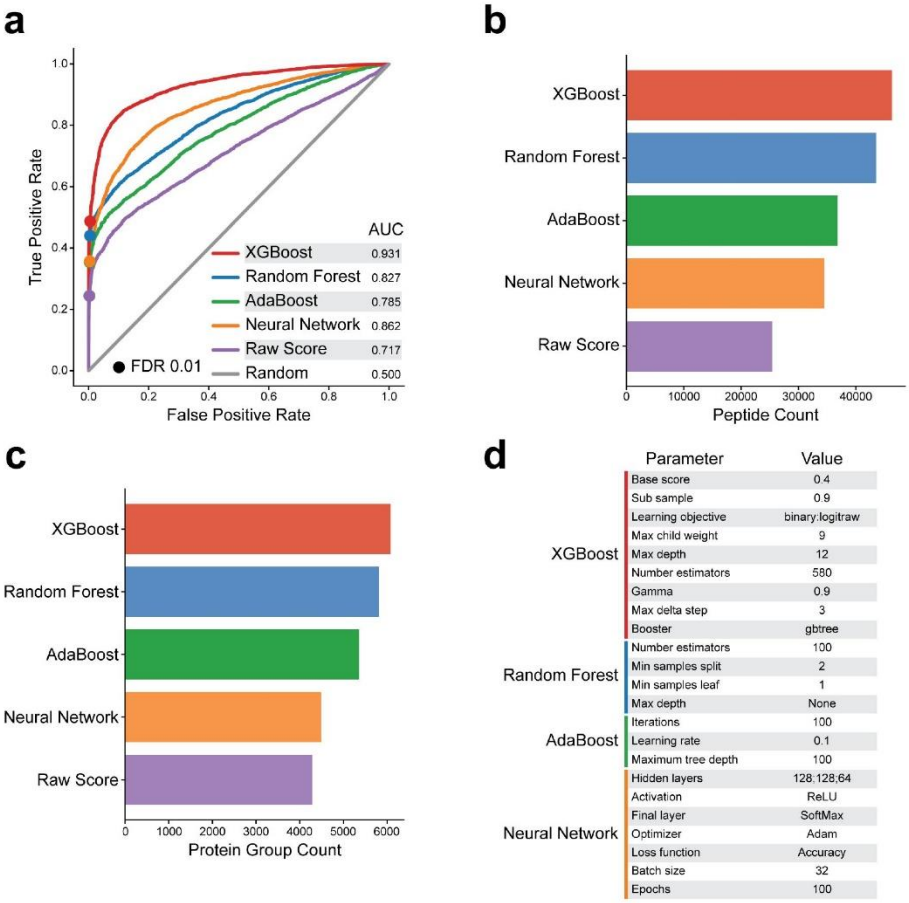
Rank	Feature	Gain
1	Precursor has isotope pattern	73.1
2	Score	42.7
3	Retention time error	13.4
4	Number of modifications	12.9
5	Number of missed cleavages	11.4
6	Fragment has isotope pattern (1)	10.0
7	Precursor mass error [ppm]	9.0
8	Charge	7.3
9	Sample fragment overlap	7.2
10	Fragment has isotope pattern (2)	6.8

e



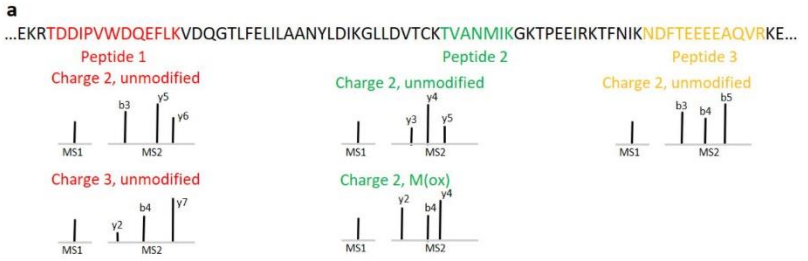
Supplementary Fig. 8: Feature space for the machine learning-based score. a, 22 'single' features for the feature matrix for calculating the machine learning score. **b,** Machine learning features derived from fragments. By default, 7 top intense fragments are considered for identification which results in a 23 + 7 * 5 = 58 dimensional feature space in total. **c,** Explanation of the fragment overlap feature. The first peptide has a fragment

overlap of 0 since the y and b ion series are not overlapping. The second peptide has overlapping y and b series and hence is its fragment overlap greater than 0. **c,** The fragment overlap score for the upper peptide is 0, since y and b ions are not overlapping and hence carry no information on the precursor mass. The lower peptide has a fragment overlap score larger than 0, since the y and b series ions are overlapping. **d,** List of the top 10 features ranked by importance according to XGBoost 'gain'. Even more important than the score is whether the precursor had an isotope pattern or is a single feature. **e,** Log-log scatter plot of feature importance according to XGBoost 'gain' for library against discovery mode. To guide the eye, we drew a straight line from the cloud of non-important features in the lower left corner to the raw score, which is expected to be of high relevance for the classification. Whether the precursor feature has an isotope pattern became much less important in the discovery mode. Features that are correlated with peptide length and charge became more important in discovery mode, presumably since the length and charge distributions of predicted spectra in the in silico library are significantly different from these distributions for peptides that are detectable in the DIA samples.



Supplementary Fig. 9: Comparison between different classification methods. We compared XGBoost, random forests, AdaBoost and fully connected multi-hidden layer neural networks to using the raw score. We tuned meta-parameters to its optimal value if applicable. **a,** ROC curves for the five classification methods. XGBoost has the highest area under the curve. **b,** Number of identified peptides when using each of the four classification Methods or the raw score in MaxDIA. XGBoost results in the highest number of peptide identifications. **c,** Number of identified protein groups when using each of the four classification Methods or the raw score in MaxDIA. XGBoost results in the highest number of peptide identifications. **d,** Optimal values of classification algorithm parameters found in grid searches.

a



b

Combination	Sequence	Charge	Modification	Ion	Sample 1	Sample 2	Sample 3	Sample 4
1	Peptide 1	2	Unmodified	Precursor	*Int 1,1*	*Int 1,2*	*Int 1,3*	*Int 1,4*
2	Peptide 1	2	Unmodified	1st fragment	*Int 2,1*	*Int 2,2*	*Int 2,3*	*Int 2,4*
3	Peptide 1	2	Unmodified	2nd fragment	*Int 3,1*	*Int 3,2*	*Int 3,3*	*Int 3,4*
4	Peptide 1	2	Unmodified	3rd fragment	*Int 4,1*	*Int 4,2*	*Int 4,3*	*Int 4,4*
5	Peptide 1	3	Unmodified	Precursor	*Int 5,1*	*Int 5,2*	*Int 5,3*	*Int 5,4*
6	Peptide 1	3	Unmodified	1st fragment	*Int 6,1*	*Int 6,2*	*Int 6,3*	*Int 6,4*
7	Peptide 1	3	Unmodified	2nd fragment	*Int 7,1*	*Int 7,2*	*Int 7,3*	*Int 7,4*
8	Peptide 1	3	Unmodified	3rd fragment	*Int 8,1*	*Int 8,2*	*Int 8,3*	*Int 8,4*
9	Peptide 2	2	Unmodified	Precursor	*Int 9,1*	*Int 9,2*	*Int 9,3*	*Int 9,4*
10	Peptide 2	2	Unmodified	1st fragment	*Int 10,1*	*Int 10,2*	*Int 10,3*	*Int 10,4*
11	Peptide 2	2	Unmodified	2nd fragment	*Int 11,1*	*Int 11,2*	*Int 11,3*	*Int 11,4*
12	Peptide 2	2	Unmodified	3rd fragment	*Int 12,1*	*Int 12,2*	*Int 12,3*	*Int 12,4*
13	Peptide 2	2	M(ox)	Precursor	*Int 13,1*	*Int 13,2*	*Int 13,3*	*Int 13,4*
14	Peptide 2	2	M(ox)	1st fragment	*Int 14,1*	*Int 14,2*	*Int 14,3*	*Int 14,4*
15	Peptide 2	2	M(ox)	2nd fragment	*Int 15,1*	*Int 15,2*	*Int 15,3*	*Int 15,4*
16	Peptide 2	2	M(ox)	3rd fragment	*Int 16,1*	*Int 16,2*	*Int 16,3*	*Int 16,4*
17	Peptide 3	2	Unmodified	Precursor	*Int 17,1*	*Int 17,2*	*Int 17,3*	*Int 17,4*
18	Peptide 3	2	Unmodified	1st fragment	*Int 18,1*	*Int 18,2*	*Int 18,3*	*Int 18,4*
19	Peptide 3	2	Unmodified	2nd fragment	*Int 19,1*	*Int 19,2*	*Int 19,3*	*Int 19,4*
20	Peptide 3	2	Unmodified	3rd fragment	*Int 20,1*	*Int 20,2*	*Int 20,3*	*Int 20,4*

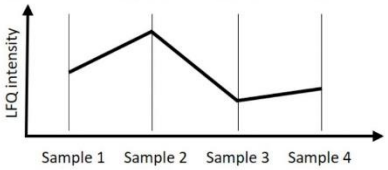
c

Sample 2	Ratio 2,1		
Sample 3	Ratio 3,1	Ratio 3,2	
Sample 4	Ratio 4,1	Ratio 4,2	Ratio 4,3
	Sample 1	Sample 2	Sample 3

d

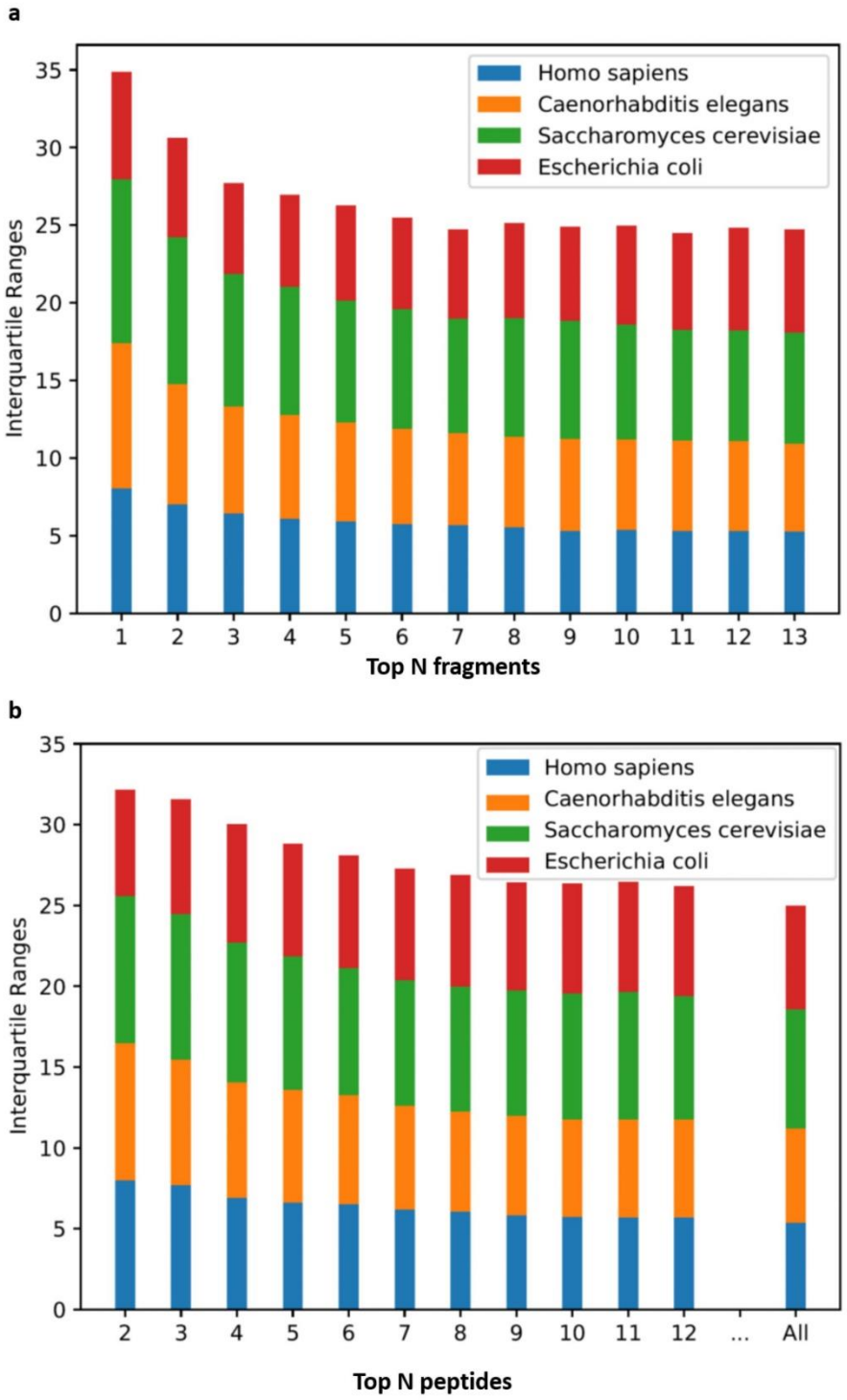
Ratio 2,1 = LFQ intensity 2 / LFQ intensity 1
Ratio 3,1 = LFQ intensity 3 / LFQ intensity 1
Ratio 4,1 = LFQ intensity 4 / LFQ intensity 1
Ratio 3,2 = LFQ intensity 3 / LFQ intensity 2
Ratio 4,2 = LFQ intensity 4 / LFQ intensity 2
Ratio 4,3 = LFQ intensity 4 / LFQ intensity 3

e

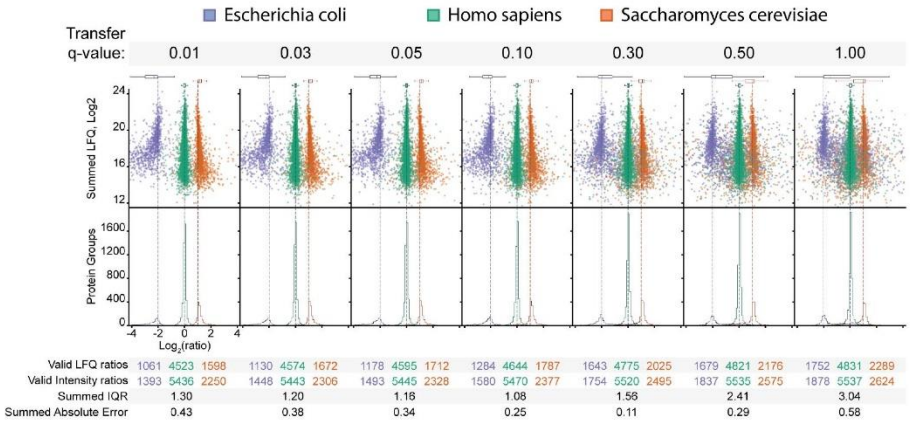


15

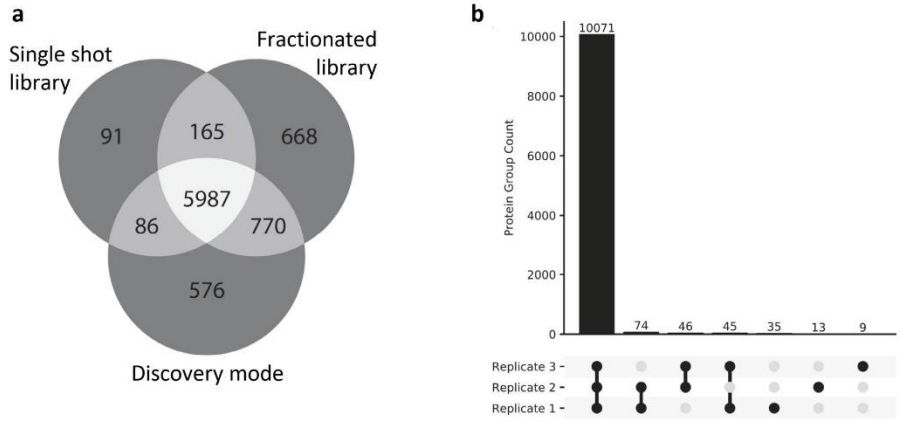
Supplementary Fig. 10: MaxLFQ algorithm for DIA. The conventional MaxLFQ algorithm for DDA consists of two parts, feature intensity normalization and protein quantification. While in the adaptation to DIA the normalization part did not change, the quantification was adapted to accommodate signals contributing from precursor and fragment features. **a,** As an example we use the protein sequence of UniProt entry P07327. Three peptides were identified, Peptide 1, unmodified with charge 2 and 3, Peptide 2, unmodified and with an oxidation of methionine, and Peptide 3, only unmodified with charge 2. These five peptide, charge and modification combinations are treated as independent intensities in the protein quantification, as was already the case in the DDA version of MaxLFQ. In DIA, also the different types of ions, precursors and fragments, are treated as separate signals. Feeding these as independent 'channels' into MaxLFQ is a natural way of implementing hybrid precursor-fragment quantification. For every combination of peptide, charge and modifications, we take the top N intense fragment peaks over the whole dataset. These N annotations are then used in every spectrum of this type for quantification. In the example we chose N = 3 for simplicity, although N is a user-definable parameter and much larger by default. (See Supplementary Fig. 11a for the influence of N on the quantification accuracy.) We also use the overall fragmentation intensity pattern of the top N fragments averaged over the whole dataset for imputation of fragments among the top N that are missing in certain samples. **b,** In the example from panel a with five peptide-charge-modification combinations and N = 3 we end up with 20 peptide-charge-modification-ion combinations. We assume that data for four samples was acquired. Then we have for this protein 20 intensity profiles over the four samples. Those intensities in this matrix which are zero we call missing, since they cannot be used for calculating ratios between samples. **c,** Next we calculate protein ratios between all pairs of samples to fill the lower triangular matrix indicated in the figure. 'Ratio 2,1' is the median of all ratios calculated from the intensities in the columns 'Sample 1' and 'Sample 2' in panel b. These are 20 if all values are present but can be less due to missing values. If the number of peptide-charge-modification combinations for which ratios can be calculated is less than the parameter 'LFQ min. ratio count' the corresponding ratio in the triangular matrix will be missing. **d,** For each ratio in panel c that is not missing we obtain one equation for the determination of the four LFQ intensities. (One for each sample.) This system of equations is usually over-determined and a least-squares best fit is obtained. **e.** Result of this operation is the profile of non-negative LFQ intensities over the four samples.


a
Homo sapiens
Caenorhabditis elegans
Saccharomyces cerevisiae
Escherichia coli
Interquartile Ranges
0
5
10
15
20
25
30
35
1
2
3
4
5
6
7
8
9
10
11
12
13
Top N fragments
b
Homo sapiens
Caenorhabditis elegans
Saccharomyces cerevisiae
Escherichia coli
Interquartile Ranges
0
5
10
15
20
25
30
35
2
3
4
5
6
7
8
9
10
11
12
...
All
Top N peptides

Supplementary Fig. 11: Optimization of number of top fragments and peptides. a, Summed inter-quartile ranges for the four-species benchmark dataset by XY et al. as a function of the number of top intense fragments used for quantification. The accuracy is increasing with rising number of fragments and plateauing around seven fragments after which no noticeable improvement happens. **b,** Same as in panel a but optimizing the number of top intense peptides used for quantification. The more peptides are taken, the higher is the quantification accuracy.



Supplementary Fig. 12: Scanning through values for the transfer q.value. We analyzed the Bruker tims-TOF pro three-species benchmark data using a range of values for the transfer q-value between 0.01 and 1. We provide summed inter-quartile ranges of species-specific ratio distributions as a measure of variability. Summed absolute errors are the deviations of the expected value for each species.



Supplementary Fig. 13: Text. a, Venn diagram of protein identifications mapped to Entrez gene identifiers for the single shot BoxCar DIA samples using three different library approaches. In particular, comparing protein identifications between fractionated library and discovery approach shows good agreement of results. **b,** Venn diagram-like comparison of replicate-specific identifications in the fractionated BoxCar DIA samples analyzed in discovery mode. Only very few protein groups were not identified in all three replicates.

Supplementary notes to 'MaxDIA enables highly sensitive and accurate library-based and library-free data-independent acquisition proteomics' by Sinitcyn et al.

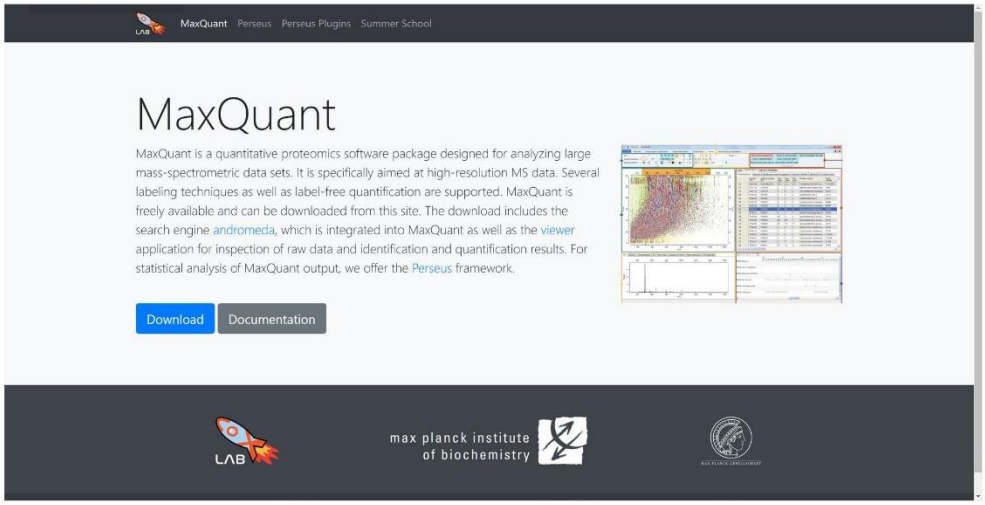
How to run MaxDIA in library mode

Summary: In order to enable MaxDIA for your DIA runs, after loading your mass spectrometry output data (raw data) into MaxQuant and setting your experiment design and the number of threads you'd like to utilize for your MaxQuant run, you can select either "Max DIA", "TIMS MaxDIA" or "BoxCar MaxDIA" from the "Type" menu within the "Group-specific parameters". Doing so will bring up a menu where you can specify your library files. These files include the peptide, evidence and msms text files from your DDA MaxQuant runs.

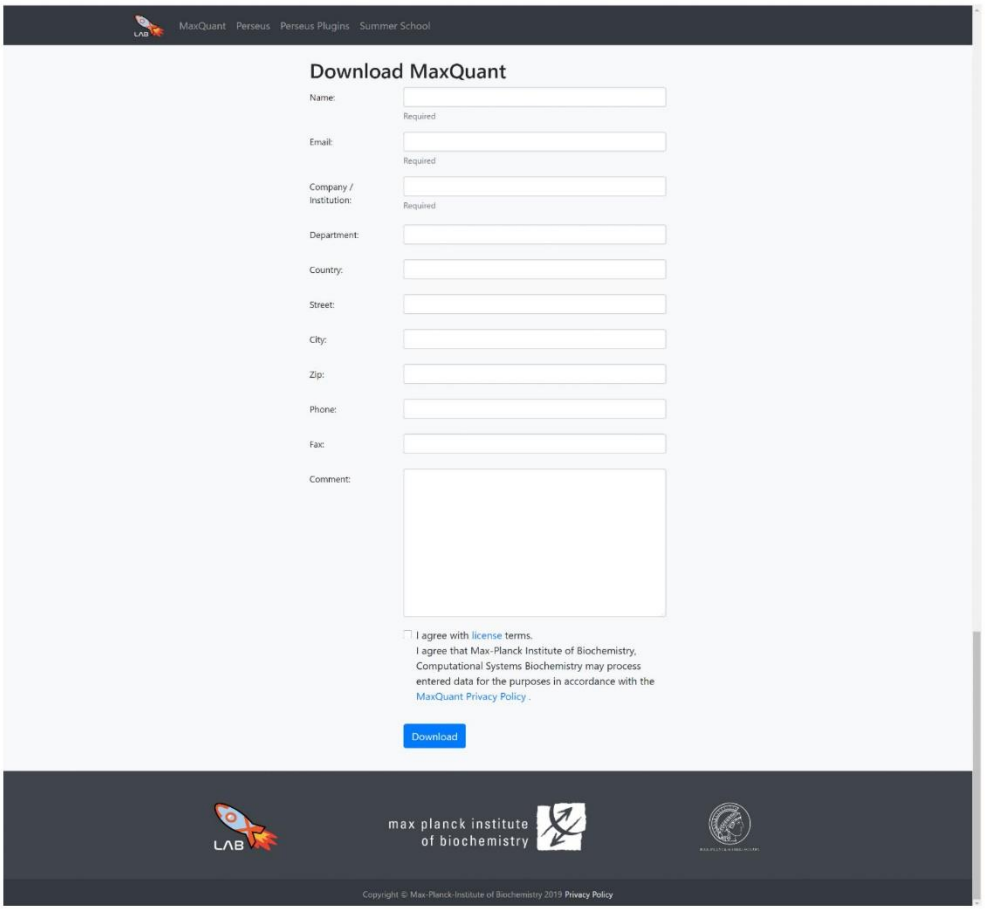
Note: To be able to run MaxQuant, .NET Core 2.1 needs to be installed. Please visit https://dotnet.microsoft.com/download/dotnet-core/2.1 and install the SDK x64."

Steps:

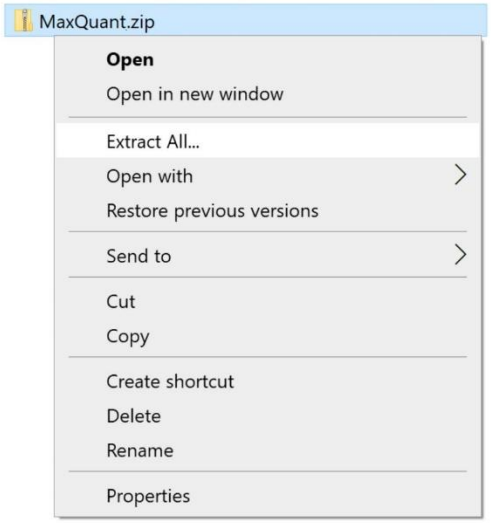
1. Using your internet browser, navigate to https://maxquant.org/



2. Click on the blue "Download" button to navigate to the download form.

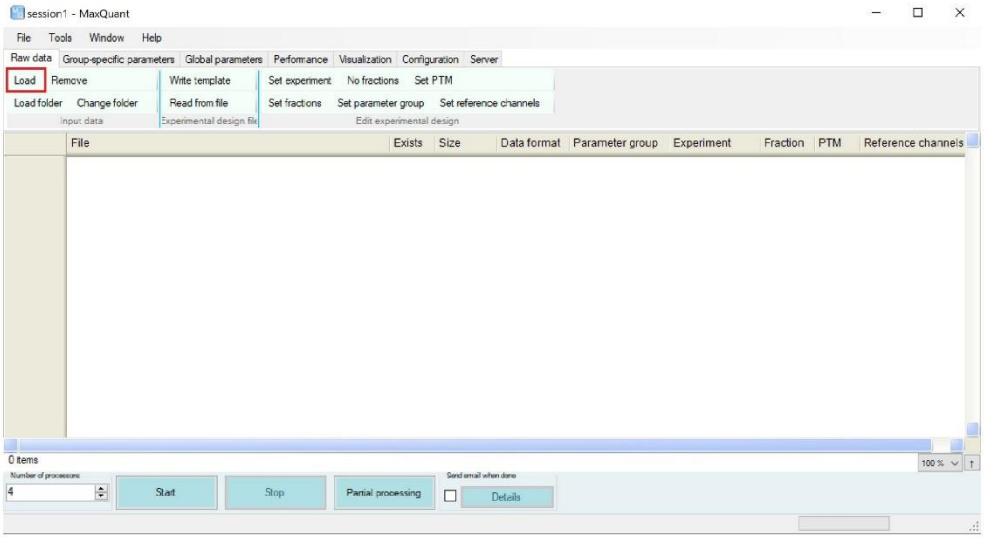


3. Fill in the form with your details and click on the check box at the end of the form to confirm your agreement with the MaxQuant license terms.
4. Click on the blue “Download” button to download MaxQuant.
5. Navigate to your downloads folder on your PC, where the zipped MaxQuant folder has been downloaded to.

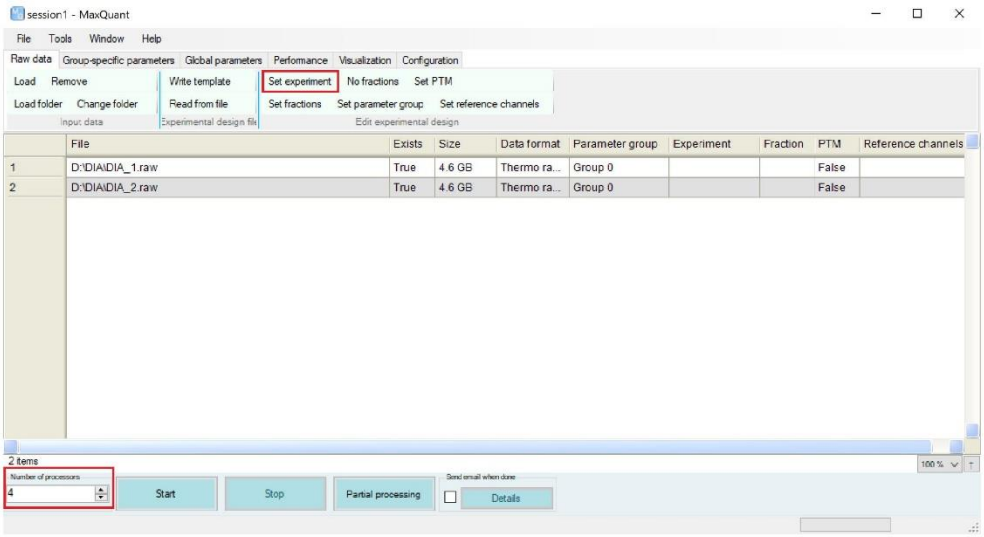


6. Extract the contents of the zipped MaxQuant folder you downloaded.

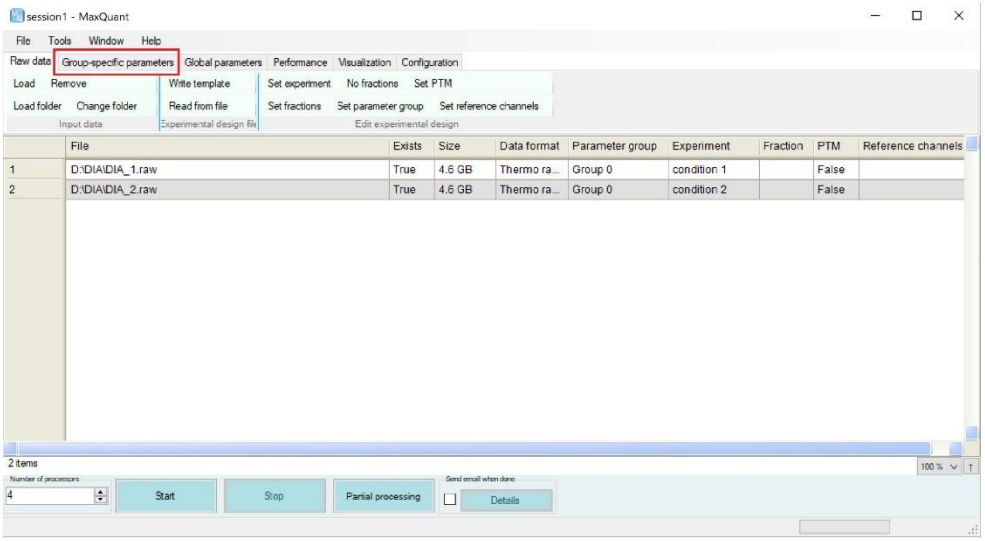
7. After extraction, open the extracted MaxQuant folder and double click on MaxQuant.exe to run MaxQuant.



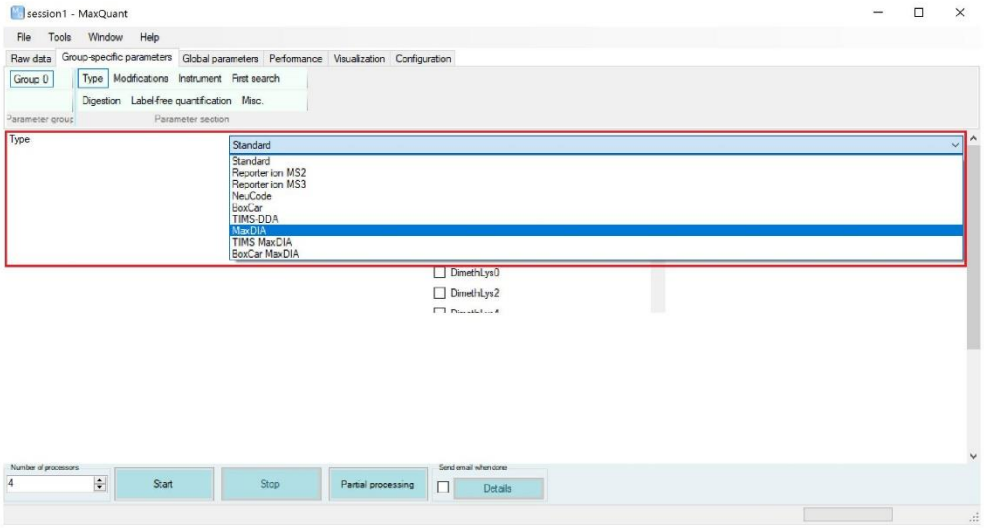
8. Click on the "Load" button to load your mass spectrometry output data (raw data) into MaxQuant.



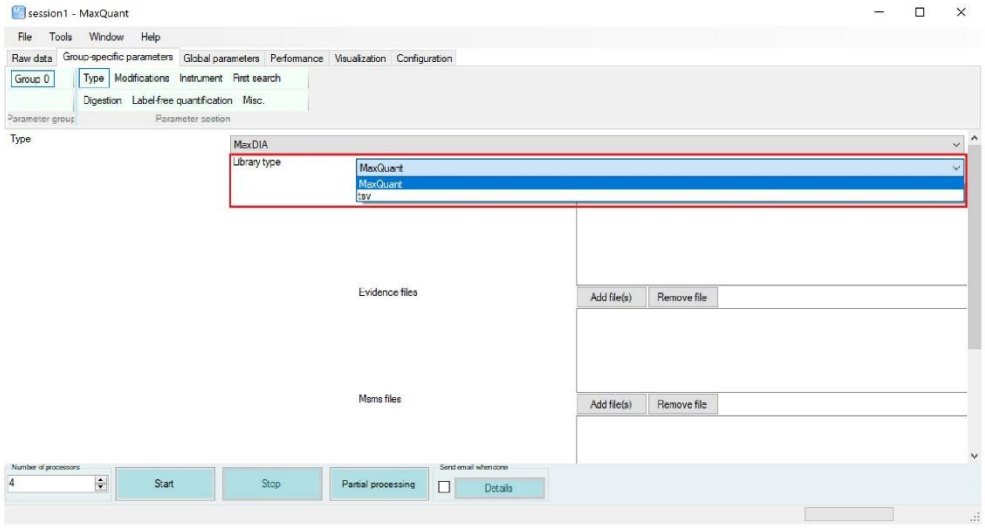
9. Now you can set the experiment design and the number of threads to be utilized by MaxQuant. Most PCs have two threads per core. You can simply press the Windows key on your PC and type "System Information", press enter and look at the number of "Logical Processors" to find out the maximum number of threads you can set. It is recommended to have at least 4 GB of Ram per utilized thread (e.g. 4 threads would need 16 GB of Ram).



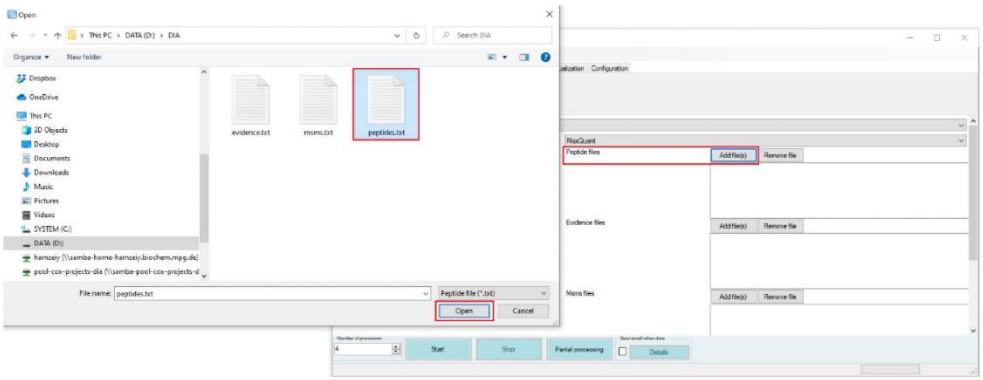
10. Next move on to the “Group-specific parameters” tab.



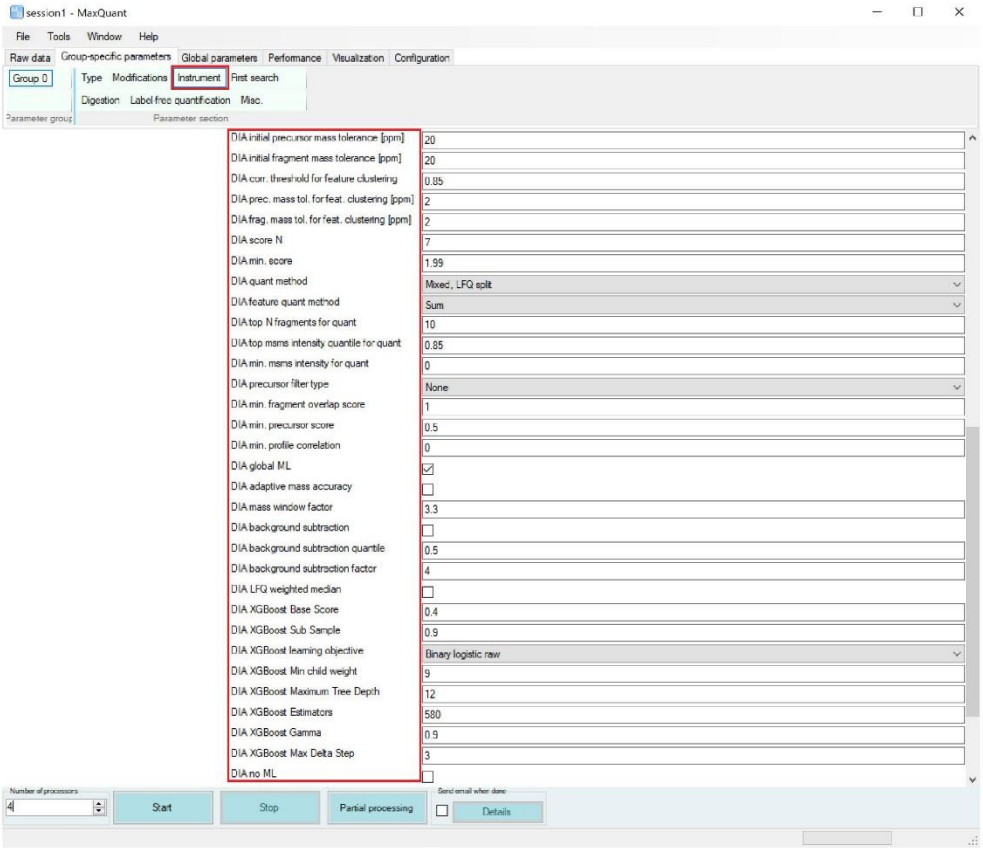
11. Here you can select the type of your mass spectrometry runs. There are three different MaxDIA algorithms available, MaxDIA, TIMS MaxDIA and BoxCar MaxDIA. Depending on your runs, choose the appropriate one.



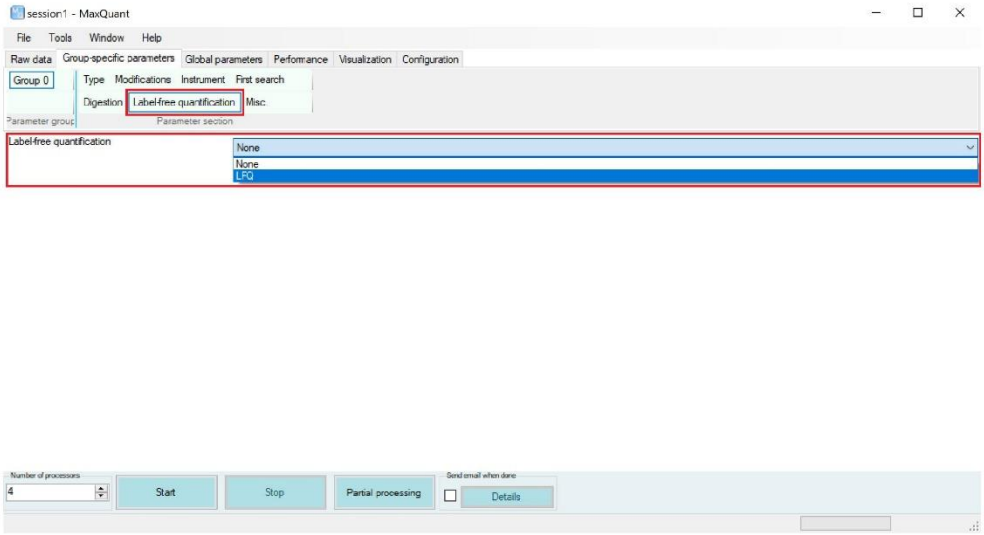
12. Next, you can choose the “Library type”. Choose “MaxQuant” for DDA library runs which have been processed with MaxQuant and “tsv” for other third party software which support a tsv output format.



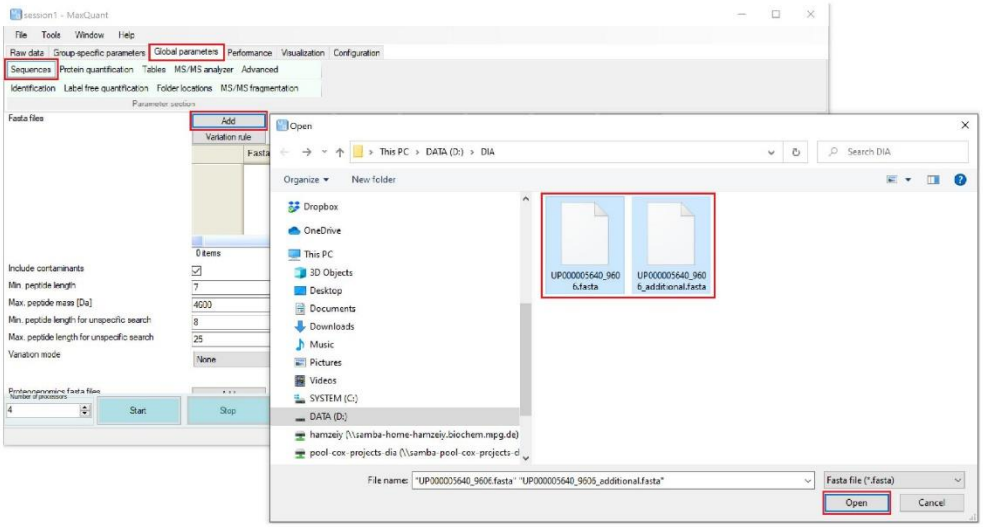
13. After choosing the library type, the library files should be added to each relevant section. The “peptides.txt”, “evidence.txt” and “msms.txt” files can be found in the “txt” folder of the “combined” folder of your DDA library runs with MaxQuant.



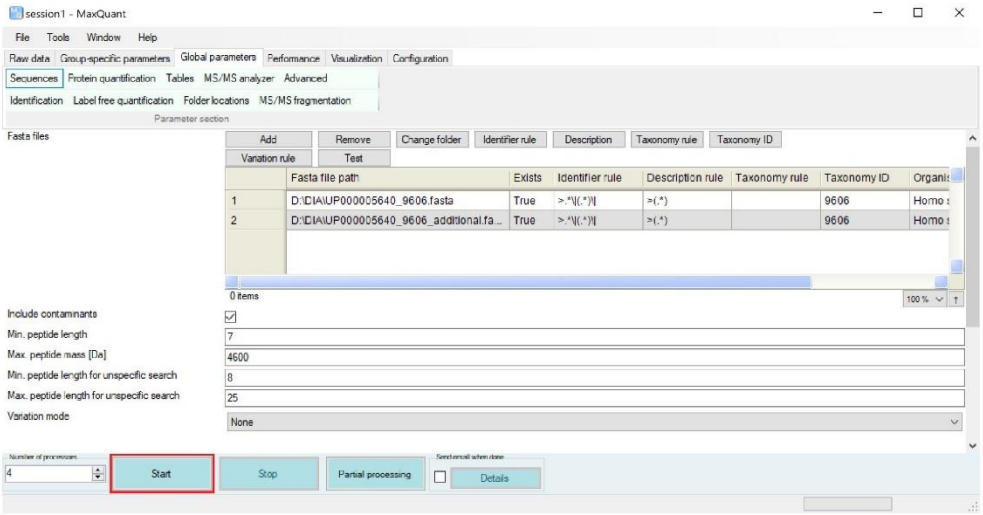
14. In the “Instrument” section, you can find many DIA related parameters. These parameters are further explained within the table at the end of this document.



15. MaxQuant's label free quantification algorithm can be used for DIA samples too. To enable this, navigate to the "Label-free quantification" section and select "LFQ" from the drop-down menu.



16. On the "Global parameters" tab, you can choose the appropriate FASTA files for your data under the "Sequences" section. You can download FASTA files for different organisms from the UniProt ftp server (ftp.uniprot.org) under: /pub/databases/uniprot/current_release/knowledgebase/reference_proteomes



17. You can now start your analysis.

How to run MaxDIA in discovery mode

Summary: Running MaxDIA is discovery mode is identical to the library mode in every step except for the library files used (step 13 of library mode). Use *in silico* generated library files to run MaxDIA in discovery mode and the relevant FASTA files. Follow the steps below to download *in silico* libraries for most common species.

Steps:

1. Navigate to http://annotations.perseus-framework.org/.

2. Click on “DiscoveryLibraries”.

3. Here you can choose your organism of choice.

4. First download the relevant FASTA files. Then depending on the number of missed cleavages choose the relevant folder.

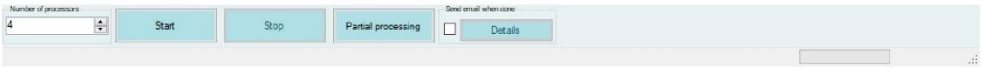
5. Here you can find the three library files needed for the discovery mode. You should unzip these files before use in MaxQuant.

How to submit results to the PRIDE repository

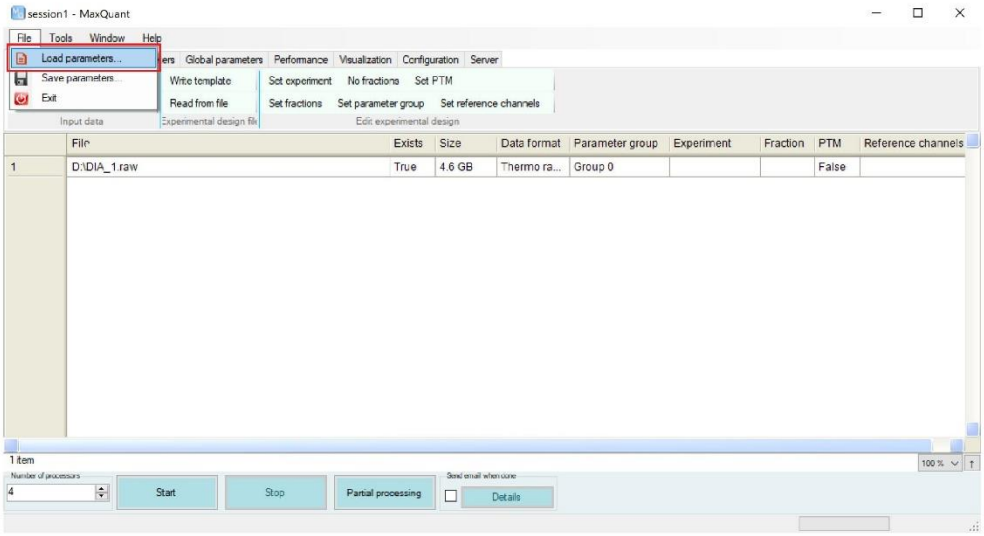
Summary: The PRIDE database has two main types of submissions “Complete Submission” and “Partial Submission”. The main different between both types of submissions is that in Complete Submissions the results (e.g. peptide and protein evidences) are provided in a standard file format such as mzTab or mzIdentML. In addition, Complete submissions received a DOI. MaxQuant supports the mzTab file format

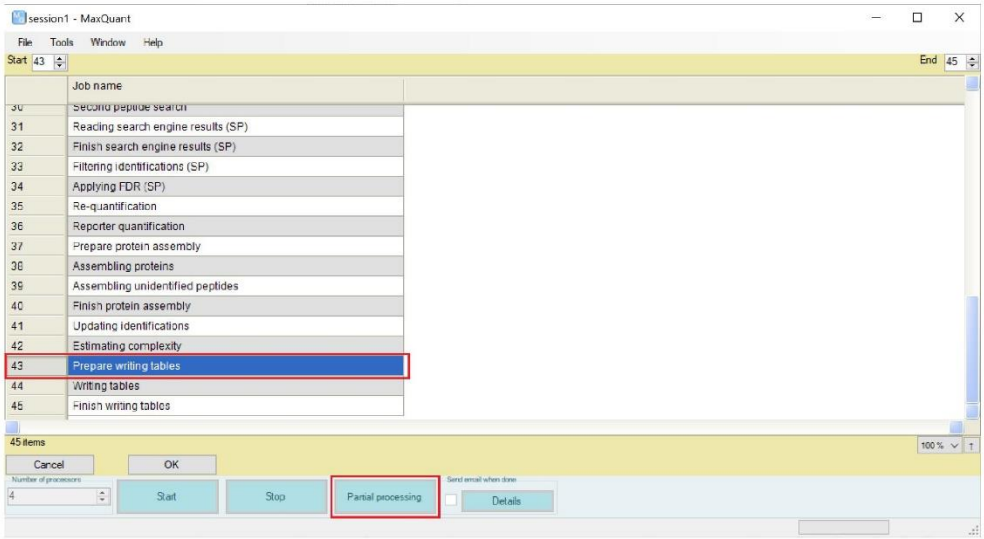
to store its results, which is needed for the PRIDE complete submission. To generate the mzTab file, simply enable it from the “Tables” menu of the “Global parameters”.

Steps:



1. To enable the mzTab output file, simple enable it from the “Tables” menu of the “Global parameters”. It is disabled by default.





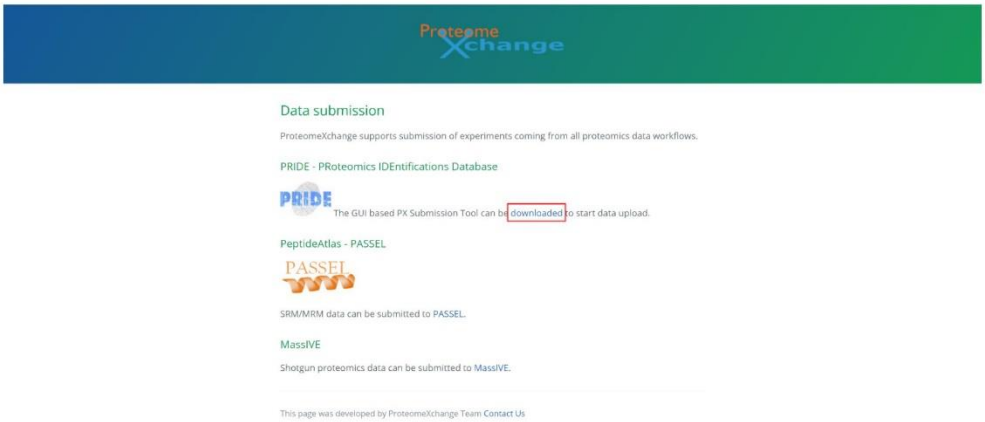
Note: You can also enable the mzTab option as described in step one and use “Partial processing” to simply only generate the mzTab file format for previously processed files by loading the relevant mqpar.xml file within the folder containing your raw mass spectrometry data.

Prepare the Pride Complete submission:

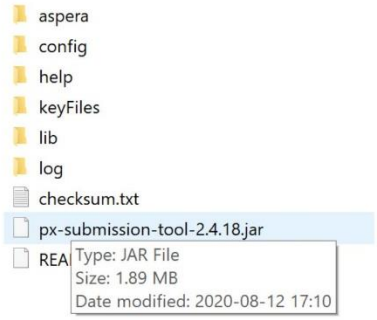
Summary: To make a complete Pride submission, you should download the submission tool from ProteomeXchange and follow the steps.

Steps:

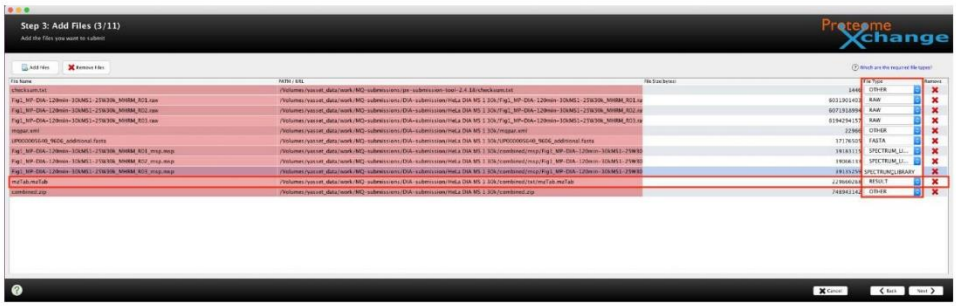
1. Navigate to http://www.proteomexchange.org/submission/index.html.



2. Download the submission tool and extract the contents of the zip file. Make sure to have java installed on your PC. The latest version of java can be downloaded and installed from https://www.java.com/en/download/.



3. Double click on the jar file or refer to the README file for instruction on running the tool from the command line. Follow the steps accordingly.
4. After adding the title, sample and protocol description in the first two panels of the ProteomeXchange submission tool, the user will arrive to a panel where files should be provided:



For MaxDIA Complete submissions the following files should be provided:

- **The mzTab File** (File Type RESULT): The mzTab contains the peptide and protein identifications in a standard file format including the references to the spectra use for the identification and the reference spectral library. The mzTab file is located in .../combined/txt/.
- **RAW files** (File Type RAW): The RAW files contain the original spectra capture by the mass spectrometer.
- **Protein FASTA database** (File Type FASTA): Protein database used in MaxDIA to map the peptides from the spectral library to the protein sequences.
- **Parameters file mqpar.xml** (File Type Other): The mqpar.xml contains all the parameters of the experiment including search parameters such as enzyme, modifications and statistical thresholds such FDRs. This file can be found where you have stored your RAW files.

- **Spectrum library references** (File Type Spectrum Library): MaxDIA generates with the mzTab a list of spectrum library files (extension MSP) which contains all the identified spectra from the original spectral library generated with the DDA data or the in-silico libraries. The MSP files are located in .../combined/msp/.
- **combined.zip** (File Type Other): In complete submissions it is important to provide also the MaxDIA combined folder in a compressed format. This folder contains additional information not included in the mzTab that are important for the users to understand the full experiment. This folder can be found where you have stored your RAW files.

Note: PRIDE recommends to perform two separate submissions for DDA and DIA data even if they are part of the same study. The user can cite or mention both accessions in the main manuscript. In this way, the DDA data used to generate the spectrum libraries can be submitted as one project and the DIA data with the resulting spectrum libraries from the DDA experiment can be submitted as a different project.

Table of all MaxDIA parameters

Parameter name (GUI)	Location in GUI Tabs	Location within GUI Tab	Parameter name (mqpar.xml)	Description
Type	Group-specific parameters	Type	lcmsRunType	This parameter can now be set to "MaxDIA", "TIMS MaxDIA" and "BoxCar MaxDIA" to turn on the MaxDIA algorithm for both library-based DIA and discovery DIA proccesing of LC-MS/MS-based proteomics runs.
Library type ("Type" must be set to "MaxDIA", "TIMS MaxDIA" or "BoxCar MaxDIA")	Group-specific parameters	Type	diaLibraryType	This parameter can be set to "MaxQuant" or "tsv", depending on the source of the library to be used for the MaxDIA algorithm
Peptide files ("Library type" must be set to "MaxQuant")	Group-specific parameters	Type	diaPeptidePaths	By clicking "Add file(s)", MaxQuant peptides.txt output file(s) or in silico peptides files in the MaxQuant output format can be defined
Evidence files ("Library type" must be set to "MaxQuant")	Group-specific parameters	Type	diaEvidencePaths	By clicking "Add file(s)", MaxQuant evidence.txt output file(s) or in silico evidence files in the MaxQuant output format can be defined
Msms files ("Library type" must be set to "MaxQuant")	Group-specific parameters	Type	diaMsmsPaths	By clicking "Add file(s)", MaxQuant msms.txt output file(s) or in silico msms files in the MaxQuant output format can be defined

Libraries ("Library type" must be set to "tsv")	Group-specific parameters	Type	diaLibraryPaths	By clicking "Add file(s)", library files in the tsv format can be defined
Min. DIA peak length	Group-specific parameters	Instrument	diaMinPeakLen	Minimum number of MS1 or MS2 scans for defining a 3D peak in DIA data
DIA initial precursor mass tolerance [ppm]	Group-specific parameters	Instrument	diaInitialPrecMassTolPpm	Indicates the mass tolerence for the initial search
DIA initial fragment mass tolerance [ppm]	Group-specific parameters	Instrument	diaInitialFragMassTolPpm	
DIA corr. threshold for feature clustering	Group-specific parameters	Instrument	diaCorrThresholdFeatureClustering	
DIA prec. mass tol. for feat. clustering [ppm]	Group-specific parameters	Instrument	diaPrecTolPpmFeatureClustering	
DIA frag. mass tol. for feat. clustering [ppm]	Group-specific parameters	Instrument	diaFragTolPpmFeatureClustering	
DIA score N	Group-specific parameters	Instrument	diaScoreN	
DIA min. score	Group-specific parameters	Instrument	diaMinScore	
DIA quant method	Group-specific parameters	Instrument	diaQuantMethod	Indicates the quantification method used for DIA data
DIA feature quant method	Group-specific parameters	Instrument	diaFeatureQuantMethod	
DIA top N fragments for quant	Group-specific parameters	Instrument	diaTopNForQuant	
DIA top msms intensity quantile for quant	Group-specific parameters	Instrument	diaTopMsmsIntensityQuantileForQuant	Indicates the top MS/MS intensity quantile to be used for quantification
DIA min. msms intensity for quant	Group-specific parameters	Instrument	diaMinMsmsIntensityForQuant	
DIA precursor filter type	Group-specific parameters	Instrument	diaPrecursorFilterType	
DIA min. fragment overlap score	Group-specific parameters	Instrument	diaMinFragmentOverlapScore	
DIA min. precursor score	Group-specific parameters	Instrument	diaMinPrecursorScore	

DIA min. profile correlation	Group-specific parameters	Instrument	diaMinProfileCorrelation	
DIA global ML	Group-specific parameters	Instrument	diaGlobalMl	Indicates whether to perform the machine learning on a per run basis or on the entire data set (global)
DIA adaptive mass accuracy	Group-specific parameters	Instrument	diaAdaptiveMassAccuracy	
DIA mass window factor	Group-specific parameters	Instrument	diaMassWindowFactor	
DIA XGBoost Base Score	Group-specific parameters	Instrument	diaXgBoostBaseScore	XGBoost base score parameter
DIA XGBoost Sub Sample	Group-specific parameters	Instrument	diaXgBoostSubSample	XGBoost sub sample parameter
DIA XGBoost learning objective	Group-specific parameters	Instrument	diaXgBoostLearningObjective	XGBoost learning objective parameter
DIA XGBoost Min child weight	Group-specific parameters	Instrument	diaXgBoostMinChildWeight	XGBoost minimum child weight parameter
DIA XGBoost Maximum Tree Depth	Group-specific parameters	Instrument	diaXgBoostMaximumTreeDepth	XGBoost maximum tree depth parameter
DIA XGBoost Estimators	Group-specific parameters	Instrument	diaXgBoostEstimators	XGBoost estimators parameter
DIA XGBoost Gamma	Group-specific parameters	Instrument	diaXgBoostGamma	XGBoost gamma parameter
DIA XGBoost Max Delta Step	Group-specific parameters	Instrument	diaXgBoostMaxDeltaStep	XGBoost maximum tree depth parameter
DIA no ML	Group-specific parameters	Instrument	diaNoMl	Parameter to turn off the machine learning

4.3 Perseus plugin 'Metis' for metabolic pathway-centered quantitative multi-omics data analysis supporting static and time-series experimental designs

Time-series omics data, available from circadian studies provide a unique opportunity to infer new knowledge on dynamic biological processes by applying multi-omics data analysis techniques. We take mice liver transcriptomics, proteomics, phosphoproteomics, metabolomics, and lipidomics circadian data and by utilizing a network-based method using a large-scale metabolic network reconstruction provided by the BioModels database, we look for enzyme activity regulation. In the context of this PhD work, data were gathered from different sources and processed in a suitable manner, all necessary design, implementation of the Perseus code-base was executed, and the following manuscript was written along with the other co-authors.

Contributions to the following correspondence within the context of this thesis include the design and implementation of the network-based multi-omics data analysis approach, data analysis and writing of the manuscript.

Hamid Hamzeiy, Daniela Ferretti, Maria S. Robles, and Jürgen Cox. "Perseus plugin 'Metis' for metabolic pathway-centered quantitative multi-omics data analysis supporting static and time-series experimental designs." Submitted to Cell Systems, 2021

Perseus plugin 'Metis' for metabolic pathway-centered quantitative multi-omics data analysis supporting static and time-series experimental designs

Hamid Hamzeiy[1], Daniela Ferretti[1], Maria S. Robles[2*], and Jürgen Cox[1, 3*]

[1]Computational Systems Biochemistry Research Group, Max Planck Institute of Biochemistry, Martinsried, Germany

[2]Institute of Medical Psychology, Faculty of Medicine, LMU, Germany

[3]Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway

*Correspondence: cox@biochem.mpg.de, charo.robles@med.uni-muenchen.de

Summary (134/150)

We introduce Metis, a new plugin for the Perseus software aimed at analyzing quantitative multi-omics data based on metabolic pathways. Data from different omics types are connected through reactions of a genome-scale metabolic pathway reconstruction. Metabolite concentrations connect through the reactants, while transcript, protein and protein posttranslational modification (PTM) data are associated through the enzymes catalyzing the reactions. Supported experimental designs include static comparative studies and time series data. As an example for the latter, we combine circadian mouse liver multi-omics data and study the contribution of cycles of phosphoproteome and metabolome to enzyme activity regulation. Our analysis resulted in 52 pairs of cycling phosphosites and metabolites connected through a reaction. The time lags between phosphorylation and metabolite peak show non-uniform behavior, indicating a major contribution of phosphorylation in the modulation of enzymatic activity.

Keywords

Multi-omics, circadian rhythms, transcriptomics, proteomics, phosphoproteomics, metabolomics, metabolic networks, enzyme activity regulation, Perseus, Metis

Introduction

Studying two or more types of biomolecules simultaneously in omics studies is of great benefit since it can reveal information that is not apparent when each of the omics dimensions is considered separately. For instance, studying transcriptome and proteome may reveal nodes of posttranscriptional regulation (Buccitelli and Selbach, 2020; Cox and Mann, 2012) which are not apparent in the transcriptomic or proteomic data alone. Another important example is expression quantitative trait loci (Cheung et al., 2005; Morloy et al., 2004) in which genetic association is correlated with gene expression to shed light on the relationship between traits and expression driven cellular processes. The combined analysis of multiple omics dimensions is challenging for multiple reasons. First of all, the quantitative measurements in each technology separately have to be of sufficiently high quality before correlations with other domains can make sense. For instance, early day studies of the relationship between cellular mRNA and protein levels found them to be nearly uncorrelated (Gygi et al., 1999), which was likely due to the shortcomings of at least one of the technologies. Since then a correlation has been confirmed in many cell types in steady state with typically moderate positive values of the Pearson correlation coefficients, despite all the differences in details between transcriptome and proteome (Wang et al., 2019). Further obstacles for multi-omics analysis are limited dynamic range and the resulting missing values problem inherent to most omics technologies. Furthermore, a statistical challenge arises, when performing all against all comparisons of variables in one technology to variables in another technology. The number of statistical tests for pair-wise correlations explodes and therefore, either a large number of false positives is created when working with p-value thresholds, or potentially meaningful truly positive signals are lost due to the necessity of stringent false discovery rate control.

Here we introduce a software solution to this problem, for the cases in which a correlative analysis involves untargeted metabolome data in combination with one or more other omics technologies, which we connect through the reactions of a metabolic pathway. For this purpose, we developed Metis, which is a plugin for the Perseus software (Tyanova et al., 2016), and which we describe in this manuscript. Perseus is a comprehensive platform for omics data analysis, which was developed with a user in mind, who is a life science

researcher but not necessarily holds a degree in bioinformatics. Hence, we expect to enable a large user base with this type of comparative multi-omics analysis, in contrast to other software tools that are targeted at programmers or bioinformatics specialists.

While Metis can be applied to any kind of experimental design, we focus here on an example with time series data highlighting an application to circadian multi-omics integration. Circadian rhythms are endogenous and self-sustainable oscillators, present in most living organisms, which drive daily cycles of molecular and metabolic processes (Finger et al., 2020). The molecular mechanism of the clock, built on transcriptional and translational feedback loops, regulates the expression of ~20% of the genes in any given tissue in mammals. Additionally, post-transcriptional and post-translational mechanisms are reported to play an essential role in circadian regulation of metabolism (Robles et al., 2014, 2017). Metis allows the investigation of cross-correlations between quantitative changes of a metabolic enzyme at different molecular level, such as transcript, protein, phosphorylation status, and the abundance changes of the products and reactants of its catalyzed reactions, aiding to pinpoint key regulatory enzymatic mechanisms. Regulatory nodes could be modulated by phosphorylation-dependent enzyme activity but also by enzyme expression changes at the protein and/or transcript level, which can be distinguished by integrating the proteome and transcriptome in the PTM data analysis. The integration of diverse temporal dynamic omics datasets together with rhythmic metabolite profiles from mouse liver uncover phosphorylation as a major enzymatic regulatory mode. Rhythmic phosphorylation of metabolic enzymes regulates its temporal activity and thus metabolic reactions across the day.

Results

Perseus software and Metis plugins

The Perseus software (Tyanova et al., 2016) is a comprehensive framework for high-dimensional omics data analysis with a focus on intuitive usability by interdisciplinary users. Through its plug-in architecture it is extensible by writing code for workflow activities in multiple programming languages, like C#, R and Python (Yu et al., 2020).

Besides statistical analysis on data matrices, the study of networks is supported as well (Rudolph and Cox, 2019). For instance, the PHOTON plug-in (Rudolph et al., 2016) can be used to analyze phosphoproteomics data in the context of protein-protein interaction networks with the aim of reconstructing kinase activities (Brüning et al., 2019).

Here we extend the network capabilities of Perseus to the specific requirements of metabolic pathways with the Metis toolbox. We use genome-scale networks of metabolic reactions to interconnect data from different omics dimensions (Figure 1). We take reactions, reactants and enzymes as nodes of the network, while edges connect the reactants and enzymes to the reactions they are taking part in. The enzyme nodes can incorporate multiple types of quantitative omics data such as proteomics, phosphoproteomics and transcriptomics data. Moreover, nodes can associate to multiple quantitative datasets of different experimental designs, comprising for instance sample group comparisons or time-series data. This allows to compare metabolic reactions across diverse datasets spanning many conditions and containing temporal data, providing thus relevant functional biological information. While we focus in this manuscript on the analysis of multi-omics time series data, Metis is more generic in terms of experimental designs and is also capable of analyzing non-time series data (Supplementary Figure 1).

The alternative to this network-based analysis of multiple omics dimensions would be an approach based on all pair-wise comparisons between the molecules in the omics datasets without filtering through a network. However, the connection of omics features through the reaction network is crucial to the analysis for the reason of statistical significance of comparisons. To illustrate this, we provide the following example: assume a comparison of untargeted metabolomics data with 1,000 compounds profiled over a time series with 10,000 phosphosites from samples of the same time series. All pair-wise comparisons between phosphosites and metabolites would amount to 10,000,000 pairs. One would then perform one statistical test per pair, for instance, to check if the Pearson or Spearman correlation is significantly different from zero. Considering in addition time-lagged correlations would further increase the number of tests on the order of ten-fold. Using a moderate p-value threshold would result in many false positives with such a high number

of hypotheses tested simultaneously. For instance, a p-value threshold of 0.01 would result in on the order of 1,000,000 false significant calls. The proper alternative to this would be false discovery rate (FDR) control, for instance with randomizations for generating the null hypothesis distribution. This would, however, likely not call any of the tests as significant due to the large background of noisy comparisons. Hence, the comparison of multiple omics through a network that provides a priori knowledge about the relationship between the omics levels is crucial for their statistical analysis.

Rhythmicity estimation of multi-omics circadian time series datasets

We decided to perform an integrative multi-omics analysis of publicly available datasets assaying *in vivo* circadian dynamics in mouse liver. In order to achieve this, we obtained and re-analyzed the most comprehensive published omics studies in transcriptomics (Hughes et al., 2009), proteomics (Robles et al., 2014), phosphoproteomics (Robles et al., 2017), metabolomics (Krishnaiah et al., 2017) and lipidomics (Adamovich et al., 2014) (see Table 1 and Figure 2). An important selection criterion was to choose studies done using the same animal housing conditions. In all studies, mice were entrained to light-dark cycles prior to being released to constant darkness, allowing us to use the same periodicity analysis in all datasets. Consequently, we use 23.6 hours as period length since this is the approximate free running period of mice driven by the internal clock. Sampling resolution differed among those studies, varying from 1h to 4h, and we kept the original time resolution of each dataset for the integrative re-analysis. Using the Perseus cycling analysis package (see Methods), we analyze rhythmicity in all omics datasets individually, by fitting the log-transformed data to a cosine curve with a period of 23.6h and calculating the FDR using 1,000 randomizations to simulate the null hypothesis of no cycling behavior. To avoid discrepancies with the findings in the published datasets we use the same FDR cut-offs as in the original publications. In addition, we use the phosphoproteomics dataset to predict kinase activity using the PHOTON method (Rudolph et al., 2016). The Perseus session for the cycling analysis plus the respective software version are provided as supplementary material.

Cycling biomolecules

The resulting cycling analysis of the five omics datasets plus the predicted kinases with daily patterns of activity is represented in Figure 3. The total number of cycling molecules has to be interpreted with caution since it is biased by the depths of the respective technologies. The fraction of cycling molecules relative to all molecules quantified by the technology (Figure 3a) is more meaningful but still not free from biases. For instance, the ability to detect statistically significant cycling profiles strongly depends on the quantitative precision of the technology, the number of time points used per cycle and the total length of the time series in relation to the period length. Furthermore, within datasets, molecules which are close to the detection limit tend to be more difficult to consider them cycling compared to highly abundant molecules.

The profiles of all cycling biomolecules are shown as heat maps in Figure 3b, which are sorted in vertical direction by their acrophase. Out of all biomolecules, metabolites show, with 58%, the largest fraction of cycling molecules. Next frequent are transcripts, of which almost one third show rhythms in abundance with acrophases uniformly distributed across the day, as described in the original publication and in other pioneering studies done with wider spaced time points (Panda et al., 2002; Storch et al., 2002; Ueda et al., 2002). In contrast, mouse liver proteins have a smaller cycling fraction, 6% of the total, sharply peaking at two main clusters, one during the day and a second one in the middle of the night (Figure 3c). The latter cluster is due to the induction of protein translation in respond to an increase in energy levels due to feeding, occurring during the night in mice as nocturnal animals. It was found (Robles et al., 2014) that when filtering the transcripts to those for which the corresponding protein is cycling, phase relations between peaking proteins and transcripts are on average compatible with the expected time lag between transcription and translation, but with strong variations between individual transcript-protein pairs.

Protein function is often regulated by post-translational modifications rather than, or in addition to, changes in protein levels. This is the case for many proteins involved in temporally regulated signaling pathways in the liver (Robles et al., 2017). Thus, it is not surprising that 26% of the phosphorylation sites, corresponding to more than 40% of liver

proteins, display daily rhythms, almost completely independent of protein abundance cycles. Similar to rhythmic proteins, cycling phosphorylation showed two distinct clusters of acrophases, but in contrast to proteins, with the majority of peaks occurring during the day or resting phase and slightly earlier than the protein cluster. The extensive regulation of phosphorylation in mouse liver implies temporal control of kinase activity. Kinase activity is, in addition to protein levels, strongly regulated by phosphorylation, and often by autophosphorylation. Thus, while among all cycling proteins in liver, there are three kinases with rhythmic changes, 55 kinases displaying cycles of phosphorylation abundance. It is however very challenging to infer kinase activity by using phosphorylation patterns on the kinases since the majority of phosphorylation sites are of unknown function (Needham et al., 2019). Taking advantage of curated kinase-substrate relationships (Hornbeck et al., 2015) we were previously able to infer a number of kinases whose activity oscillates across the day (Robles et al., 2017). Since this prediction method is biased towards well-known kinases, we here use an alternative method to predict cycles of kinase activity, the PHOTON algorithm (Rudolph et al., 2016), which is based on the statistical analysis of protein-protein interaction networks. Applying it to the phosphoproteome data, we were able to predict 33 distinct kinases with changes in their activities across the day, corresponding to 20% of all kinases in the PHOTON analysis. Interestingly, predicted kinase activities are enriched in two temporal regions slightly preceding the phosphorylation clusters on average, indicative of a time lag between peak kinase activity and maximum substrate phosphorylation. Overall, our data shows that kinetics of molecular reactions, such as phosphorylation, can be studied using large scale time series data.

Circadian clocks and metabolism crosstalk bi-directionally. While tissue clocks regulate local metabolism, the metabolic state feeds back to the molecular clock (Brown, 2016). Accordingly, in mouse liver, which is one of the most studied organs, a large proportion of metabolites have been described to display rhythms across the day. Perseus cycling analysis of the metabolomics data resulted in more than 50% of metabolites with daily cycles, similar to what was reported in the original study (Krishnaiah et al., 2017). Rhythmic metabolites peak at diverse times of the day, many of them during the inactive

phase (Figure 3). Similarly, Perseus cycling analysis of the lipidomics data yielded cycling lipids, 7% of the total, peaking predominantly during the day as previously reported (Adamovich et al., 2014).

Reaction-based multi-omics filtering

Metabolic networks are represented within Perseus with three different node types, namely reaction, enzyme and metabolite (Figure 4). Metabolite nodes are connected to the reactions they participate in, and are classified as a substrate or a product of the reaction in question. Since enzymes are not consumed or produced via the reaction but only catalyze it, these nodes are connected to the reaction nodes in an undirected manner. All node types can be annotated within the nodes table of the network in Perseus with both qualitative and quantitative information, e.g. in the case of a reaction this can be its reversibility or rate. Edges can also be annotated with various qualitative and quantitative information. When filtering for reactions of interest, Perseus can retain all reaction nodes where a condition is either true or false and/or a certain threshold is applied to the numerical annotation at the metabolite or enzyme nodes of the metabolic network. In other words, all nodes that meet the condition/s of the filter plus the reaction nodes directly connected to them can either be retained or removed in order to reach the desired network for further analysis in the next processing step. The user also has the option to apply several filters of a certain type on different properties of the nodes and edges in union or as intersect, depending on the nature of the question. Here, we apply a filter on the q-values of the periodicity analysis of each of the three omics dimensions, namely proteomics (q-value >=0.33), phosphoproteomics (q-value < 0.1) and metabolomics (q-value <0.05), to retain all reactions that have nodes with our criteria for further analysis. Following this filtering, we looked for reactions mediated by enzymes that are not cycling at the protein level, harbor cycling phosphorylations and have rhythmic substrates and/or products. For more details on how to perform the filtering see Methods and Figure 4.

Phosphorylation as driver of dynamic enzymatic reactions

The multi-omics network analysis with Metis resulted in a metabolic sub-network with cycling metabolites of reactions mediated by enzymes with rhythmic phosphorylation

changes (Figure 5 and Supplementary Table 1). Together, those rhythmic reactions covered several metabolic processes that involve important metabolites such as NADH, AMP, CoA and amino acids. Overall we find that phosphorylation works as a regulatory switch for enzymes in key metabolic reactions in mouse liver. While phosphorylation can serve as an activating or repressive modulator, most of the time, the functional relevance of a phosphorylated residue identified in a large scale phosphoproteomics study is unknown (Needham et al., 2019). This is the case for many of the rhythmic phosphorylations in metabolic enzymes of our network as well. We thus seek to infer the functional role of the phosphorylation in the activity of some of these enzymes based on the quantitative correlation with the substrate and/or product of its reaction. An evident example of this is the case of carbamoyl-phosphate synthetase 2 (CAD), which is confirmed by our analysis, where we reproduce the fact that the enzyme is allosterically regulated by phosphorylation at S1859 (Robitaille et al., 2013) (Supplementary Figure 2).

Moreover, S26 phosphorylation of acyl-coenzyme A oxidase 1 (ACOX1), the enzyme catalyzing the first step of peroxisomal very-long-chain fatty acid oxidation, cycles with a peak in the resting phase concomitant with the nadir of FAD, the cofactor of this reaction (Figure 5a, left panel). Low cofactor levels could indicate high enzymatic activity plausibly driven by phosphorylation, leading to a temporal regulation of peroxisomal fatty acid oxidation with a peak in the inactive phase as reported for mitochondria fatty acid oxidation (Neufeld-Cohen et al., 2016). FAD is also a cofactor in first step of proline catabolism mediated by proline dehydrogenase (PRODH). In addition to the cofactor FAD, proline as substrate of this reaction is also rhythmic with a nadir during the inactive phase similar to FAD and also to the cycling profile of PRODH phosphorylation of S32. Thus, under nutrient stress during the resting phase when mice are not eating, PRODH activation would mediate proline catabolism to maintain the cellular energy levels (Pandhare et al., 2009). In contrast to what we observed for ACOX1, increased S32 phosphorylation of PRODH would lead to enzymatic inhibition and accumulation of proline during the active phase when nutrient levels are high due to food intake (Figure 5a, middle and right panels).

Another example is the crosstalk between Acetyl-CoA Synthase (ACSS2) rhythmic phosphorylation and the cycle of its enzymatic cofactor, coenzyme A (CoA). ACSS2 is rhythmically phosphorylated in S267 while S30, S263, S267 and S263 phosphorylations do not cycle. Acetyl-CoA is produced by ACSS2 using citrate and CoA as substrates, therefore the fact that the CoA cycles in antiphase to the ACSS2 S267 phosphorylation suggests that ACSS2 activity is promoted by S267 phosphorylation (Figure 5f). A similar relationship can be inferred for the CoA Synthase, COASY, which cycling phosphorylation at S177 and S182 occurred parallel to the rhythmic levels of its enzymatic product CoA (Figure 5f).

Another very interesting cross-correlation between rhythmic enzymatic phosphorylation and cycles of substrate and product metabolites is the reaction mediated by the Glycine N-methyltransferase (GNMT). GNMT catalyzes the synthesis of N-methylglycine (sarcosine) from glycine using S-adenosylmethionine (SAM) as the methyl donor, producing S-adenosylmethionine (SAM). The peak of GNMT phosphorylation at S10 in the middle of the night, likely driven by feeding as previously reported, is concomitant with the maximum levels SAH and nadir of SAM, product and substrate of its enzymatic reaction, respectively (Figure 5d). In this manner, the nutrient dependent regulation of GMNT activity via phosphorylation would impact methionine metabolism and the methyl cycle by controlling the SAM/SAH ratio. Since SAM is the methyl donor for almost all cellular methylation reactions, temporal control of GNMT activity across the day would likely contribute to the daily rhythms of RNA and histone methylation and their crosstalk to the molecular clock (Fustin et al., 2013, 2020; Greco et al., 2020).

Phase relations between metabolite concentrations and enzyme phosphorylation

In the previous section we looked into specific examples of pairs of cycling metabolites and cycling enzyme phosphorylation that passed our filters targeted at finding enzyme regulation by phosphorylation. In Figure 6 we provide a histogram of phase differences containing all such pairs found by the network analysis. The phase differences are grouped into 3h bins. The bin at 0 contains those cases for which the enzyme phosphorylation is in phase with the metabolite levels (phase difference between -1.5h and +1.5h), while the bin

at 12 contains those cases for which metabolite and phosphorylation levels are in anti-phase (phase difference between 10.5h and 13.5h). The highest bin is the one in exact anti-phase at 12. In this bin, the metabolite concentration is high when the fraction of enzymes that are phosphorylated at the respective site is low and vice versa. One needs to distinguish cases in which the metabolite is counted as a product or as a substrate of the reaction. In case the metabolite is a product, these can be interpreted as potential cases of enzyme activity repression by phosphorylation. In case the metabolite is a substrate, the interpretation is enzyme activation, since the more substrate is consumed, the greater enzyme function driven by higher phosphorylation level. Accuracy of the data, simplicity of the fit model and the binning of phases, all give some leeway to the phase relationship which can accommodate 1-2h time lags due to accumulation or consumption times of metabolite concentrations.

Another interesting region of the histogram is around time lag 0 (bins -3h, 0h, 3), which also has an increased number of cases compared to the average. Here, phosphorylation levels are close in phase with the metabolite concentration, leading to the opposite interpretation as in the 12h bin. Here the cases with substrates are interpreted as enzyme suppression by phosphorylation while cases with products are interpreted as enzyme activation. Time lags due to accumulation/consumption effects seem to be larger here as is manifested by the larger spread of phase differences onto the -3h and 3h bins.

In summary, the distribution of phase differences between metabolites and corresponding enzyme phosphorylation is non-uniform and indicative of enzyme activity regulation due to metabolite and phosphorylation crosstalk. Several of these enzyme-metabolite pairs could be rationalized in the previous subsection.

Discussion

We introduced Metis, a Perseus plugin for the joint analysis of multiple omics datasets through metabolic networks. On circadian multi-omics data for mouse liver we find enzyme phosphorylation-metabolite pairs co-occurring in reactions, that show phase

relationships indicative of activation and repression. The circadian multi-omics analysis using Metis highlights phosphorylation as a major regulatory switch of enzymatic activity regulating daily metabolic reactions in mouse liver as already shown for receptor downstream signaling pathways in this same organ (Robles et al., 2017).

We see data quality in terms of completeness of quantification over whole time series, as well as quantification accuracy as limitations to the data analysis, in particular for the metabolome and phosphoproteome data. The noisiness led us to perform as a first step circadian analysis separately in each of the omics levels and do the network-based analysis with the resulting fit parameters. With advances in data quality it will be possible, as well as of interest, to perform the network analysis with correlations across omics dimensions on the raw time profiles. We see Perseus and Metis as a very suitable framework for this purpose.

Analysis of circadian multi-omics dataset using Metis highlighted a predominant role for phosphorylation regulating the activity of metabolic enzymes and consequently metabolism. Metabolic state crosstalk to the molecular clock to thus ensure proper circadian response to metabolic changes (Brown, 2016). One mechanism of crosstalk could involve metabolic enzymes that directly modulate circadian transcriptional control by physically interacting with chromatin remodeling systems, reprograming gene expression in response to the metabolic state (Boon et al., 2020; Li et al., 2018). This reprogramming would be largely based on posttranslational mechanism leading to modifications of histones and nonhistone proteins to ultimately control their activity. Our analysis of metabolite and enzymatic activity supports this notion and even exposes complementary metabolic reactions which can impact transcription. For example, while ACSS2 phospho-dependent peak of activity in the middle of the night would promote histone acetylation (Mews et al., 2017) and transcriptional activity by generating Acetyl-CoA, food-driven phosphorylation and activation of GNMT in the night would inhibit histone methylation and thus transcriptional repression, by reducing SAM levels (Fustin et al., 2013, 2020; Greco et al., 2020). Thus, rhythms of metabolic enzymatic activity and corresponding

metabolite levels would specifically impact the circadian molecular machinery at the chromatin to ultimately entrain the molecular clock to metabolic and nutrient state.

Acknowledgements

This project was partially funded by the German Ministry for Science and Education (BMBF) funding action MSCoreSys, reference number FKZ 031L0214D. MSR is funded by the DFG (Project 329628492 – SFB 1321 and 428041612). We thank Dr. Jan Daniel Rudolph for help with PHOTON.

Author contributions

H.H. and J.C. designed the tools used within the Perseus software and H.H. and D.F. implemented them. H.H., M.S.R. and J.C. performed the analyses. M.S.R. and J.C. directed the research. H.H., M.S.R. and J.C. wrote the manuscript.

Declaration of interests

The authors declare no competing financial interests.

STAR methods

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jürgen Cox, cox@biochem.mpg.de.

Materials availability

This work consists purely of data and data analysis software for the generation of the results and thus no further materials were used.

Data and code availability

Data analysis was performed with Perseus version 1.6.15 containing the Metis plugin. All Perseus sessions and further data are available in the Mendeley dataset http://dx.doi.org/10.17632/n4nx6x999v.1,

Method details

Transcriptomics data

The mouse liver circadian transcriptomics data contained 18,647 transcripts, quantified every hour (48 time points). The series matrix was downloaded as a .txt file from the Gene Expression Omnibus (GEO) with the ID GSE11923 (Hughes et al., 2009). The header of the downloaded file was removed, only keeping the "!Sample_title". This file was then uploaded to Perseus using the "Generic matrix upload" function where all columns containing expression values were uploaded as "Main" and the "ID_REF" column was uploaded as "Text". The "ID_REF" column was used to annotate the transcripts with UniProt IDs using a Perseus annotation file shipped with the software (which was also uploaded to Perseus using the "Generic matrix upload" function) and the "Matching rows by name" function of Perseus. Rows which were not annotated with a UniProt ID were removed using the "Filter rows based on text column" function. The rows with the same UniProt ID were combined, taking the median using the "Unique rows" function. Having checked the distribution of the data using the "Histogram" function, the data was transformed by log10(x) using the "Transform" function (see Perseus session file named "Transcriptomics.sps").

Proteomics data

The mouse liver circadian proteomics data contained 3,132 proteins, quantified every three hours (16 time points). The data was obtained from the supplementary material of the original article (Robles et al., 2014) as an Excel document. The sheet named "A- Total dataset" was saved as a .txt file and uploaded to Perseus using the "Generic matrix upload" function where all columns containing expression values were uploaded as "Main" and the UniProt IDs of the proteins as "Text". Having checked the distribution of the data using

the "Histogram" function, and since the original paper reports the expression values to have been transformed by log2(x), the data was first transformed by $2^{(x)}$ and later by log10(x) using the "Transform" function. This is done so that all data sets are treated exactly in the same manner within Perseus (see Perseus session file named "Proteomics.sps").

Phosphoproteomics data

The mouse liver phosphoproteomics data contained 7,986 phosphosites quantified every three hours (16 time points). The data was obtained from the supplementary material of the original article (Robles et al., 2017) as an Excel document. The sheet named "A- Total dataset" was saved as a .txt file and uploaded to Perseus using the "Generic matrix upload" function where all columns containing expression values were uploaded as "Main", the UniProt IDs of the proteins as "Text", the phosphorylated amino acid as "Text" and the position of the phosphorylated amino acid within the protein as "Text". Since the data was already log10(x) transformed, no further processing was carried out (see Perseus session file named "Phosphoproteomics.sps").

Metabolomics data

The mouse liver circadian metabolomics data contained 224 metabolites, quantified every hour (48 time points). The data was obtained from the supplementary material of the original article (Krishnaiah et al., 2017) as an Excel document. The sheet named "Liver_data" was saved as a .txt file. For the measured metabolites we could retrieve 200 ChEBI IDs which were later used to map and annotate the metabolites according to the mouse metabolic network from the BioModels database. The ChEBI IDs were retrieved via a simple script from HMDB website using the supplied HMDB IDs within the original supplementary file provided by Krishnaiah et al., 2017. For metabolites missing HMDB IDs, metabolite names were used for manual retrieval of ChEBI IDs. The resulting .txt file was then uploaded to Perseus using the "Generic matrix upload" function where all columns containing the metabolite quantification values were uploaded as "Main". Having checked the distribution of the data using the "Histogram" function, the data was transformed by log10(x) using the "Transform" function (see Perseus session file named "Metabolomics.sps").

Lipidomics data

The lipidomics data contained 159 lipids, quantified every four hours (6 time points). The data was obtained from the supplementary material of the original article (Adamovich et al., 2014) as an Excel document. The sheet named "A." was modified to have the lipid types as a column instead of row separators and saved as a .txt file and uploaded to Perseus using the "Generic matrix upload" function where all columns containing the lipid quantification values were uploaded as "Main", the mass as "Numeric" and the type and name as "Text". Having checked the distribution of the data using the "Histogram" function, the data was transformed by log10(x) using the "Transform" function (see Perseus session file named "Lipidomics.sps").

Kinase activity prediction

Kinase activity prediction was performed using the PHOTON plugin of Perseus. The phosphoproteomics data was annotated based on the UniProt IDs with Gene IDs and ENSP IDs using the Perseus "Add annotation" function. Three different protein-protein interaction networks were used, namely BioGRID, IntAct and STRING. These were downloaded from the respective web sources and are available as supplementary data within this paper. After the analysis described below for each interaction network, the periodicity analysis is performed as explained in the Periodicity Analysis section and the resulting matrices are merged, annotated using the "Add annotation" function with "Keywords" which were later filters for "Kinase" using the "Filter rows based on categorical column" to keep only the kinases from the PHOTON prediction (see Perseus session file named "Kinase Activity Prediction.sps").

The BioGRID network .txt file and uploaded to Perseus using the "Generic matrix upload" function with the confidence column as "Numeric" and the source and target columns as "Text". The resulting matrix was converted to the Perseus "Network collection" data type using the "From matrix" function of Perseus choosing the correct "Source" and "Target" columns. Then the node degrees were calculated using the "Node degrees" function of Perseus and filtered using the "Filter nodes by numerical column" function for nodes with less than 600 degrees in order to discard nodes that are connected to too many other nodes

which would cause significant noise within the PHOTON analysis. The remaining nodes within the network were then annotated with the phosphoproteomics quantitative time series data (all the "Main" columns) using the "Annotate nodes" function based on the "Node" column of the network and the "GeneID" column of the phosphoproteomics data, selecting "Keep separate" for the "Combine copied main values" option. PHOTON analysis was done using the PHOTON plugin of Perseus where all the columns contacting the quantitative data are selected and the "Side" is selected as "twosided" and a python.exe path is given as "Executable". As output, PHOTON provides two network collections and a matrix. The matrix was further processed using the "To base identifiers" function based on the "Node" column containing the Gene IDs to retrieve the UniProt IDs for the PHOTON prediction. The columns were also renamed using the "Rename columns" function and sorted using the "Sort columns" function. The resulting matrix was saved as a .txt file using the "Generic matrix export" function.

The IntAct network .txt file and uploaded to Perseus using the "Generic matrix upload" function with the confidence column as "Numeric" and the source and target columns as "Text". The resulting matrix was filtered using the "Filter nodes by numerical column" function for the protein-protein interactions with confidence values greater than 0.5. The resulting converted to the Perseus "Network collection" data type using the "From matrix" function of Perseus choosing the correct "Source" and "Target" columns. The nodes within the network were then annotated with the phosphoproteomics quantitative time series data (all the "Main" columns) using the "Annotate nodes" function based on the "Node" column of the network and the column containing the UniProt IDs of the phosphoproteomics data, selecting "Keep separate" for the "Combine copied main values" option. PHOTON analysis was done using the PHOTON plugin of Perseus where all the columns contacting the quantitative data are selected and the "Side" is selected as "twosided" and a python.exe path is given as "Executable". As output, PHOTON provides two network collections and a matrix. The matrix was further processed to rename the columns using the "Rename columns" function and sort them using the "Sort columns" function. The resulting matrix was saved as a .txt file using the "Generic matrix export" function.

The STRING network .txt file was uploaded to Perseus using the "Raw upload" function with the "Split into columns" selected along with the "Separator" as "Tab". The resulting matrix's score column type was converted from string to numerical using the "Change column type" function. Later the score column was used to filter for interactions with a score higher than 900 using the "Filter nodes by numerical column" function. The "mode" column was also used to filter for protein-protein interactions which were categorized as "binding" using the "Filter rows based on text column" function with the search string set as "binding", without matching for case but matching for the whole word and the "Mode" selected as "Keep matching rows" and the filter mode as "Reduce matrix". The resulting matrix was further processed using the "Process text column" function to remove "10090." from the beginning of the ENS IDs within the STRING network "item_id_a" and "item_id_b" columns with the regular expression "10090\.(.*)" and no replacement string. The "Rename columns" function was used to rename the columns within the matrix containing the ENS IDs to "protein1" and "protein2" for the interacting proteins. The "Transform" function was then used on the confidence column with the transformation formula as "x/1000". Later, using the "Reorder/remove columns" function, only the columns "Confidence", "protein1" and "protein2" were kept for creating the Perseus network collection using the "From matrix" function to be used for PHOTON analysis. Prior to the PHOTON analysis, the node degrees were calculated using the "Node degrees" function which was used to filter the network for nodes with degrees less than 1000 in order to discard nodes that are connected to too many other nodes which would cause significant noise within the PHOTON analysis. The nodes within the network were then annotated with the phosphoproteomics quantitative time series data (all the "Main" columns) using the "Annotate nodes" function based on the "Node" column of the network and the column containing the ENSP IDs of the phosphoproteomics data, selecting "Keep separate" for the "Combine copied main values" option. PHOTON analysis was done using the PHOTON plugin of Perseus where all the columns contacting the quantitative data are selected and the "Side" is selected as "greater" and a python.exe path is given as "Executable". As output, PHOTON provides two network collections and a matrix. The matrix was further processed to rename the columns using the "Rename columns" function

and sort them using the "Sort columns" function. The resulting matrix was saved as a .txt file using the "Generic matrix export" function.

Periodicity analysis

All the data acquired from previous publications and the kinase activity prediction were analyzed using Perseus's time-series analysis toolkit. Perseus performs this analysis in a permutation-based FDR-controlled manner and calculates the amplitude of the change and the peaking time for each case by fitting the data to a cosine function (Tyanova et al., 2016). The results were then filtered to define "cycling" and "non-cycling" entries according to the original q-values recommended by each of the publications that the data originated from and q-value less than 0.1 for the kinase activity prediction data using the "Filter rows based on numerical/main column" function. The data was annotated using the "Categorical annotation rows" and "Numerical annotation rows" functions for all the measurements within the 48 hours. The last five time points of the transcriptomics data were removed due to systematic abnormalities observed in the data using the "Reorder/remove columns" function. Using the numerical annotation, the circadian analysis was done using the "Periodicity analysis" function with the period set to 23.6, FDR set to 1 and number of randomizations set to 1000. The heat maps were made by collapsing and averaging the measurements to 3 hour intervals using "Average groups" function and the zero and 24 time point were calculated using the same data points. Prior to visualizing the heat map the "Z-score" function was used for normalization and the "Hierarchical clustering" function was used without the row and column trees (see Perseus session files named "Lipidomics.sps", "Metabolomics.sps", "Phosphoproteomics.sps", "Proteomics.sps", "Transcriptomics.sps" and "Kinase Activity Prediction Cycling Analysis.sps").

Whole genome metabolic networks

The metabolic network used in this study and available at http://annotations.perseus-framework.org/ within the "MetabolicNetworks" folder for 11 most popular model organisms are based on data downloaded from the BioModels database (Path2Models)

(Büchel et al., 2013). These metabolic networks are parsed upon retrieval from the BioModels database as an .xml file and reduced to two text files containing the edges and the node annotations of the network. The .txt file with the edges of the network contains two columns, "Source" and "Target", while the .txt file with the node annotations contains two columns, "Node" and "Type". These files can simply be uploaded to Perseus as explained in the following section and used.

Network mapping and filtering

The network .txt files for mouse were uploaded to Perseus using the "Generic matrix upload" function. For the edge table both the "Source" and "Target" columns are selected as "Text" and for the node annotation table the "Node" column is selected as "Text" and the "Type" column as "Categorical". Using the "From matrix" function, the matrix containing the edges of the network was converted to a Perseus network collection. Later using the matrix containing the node annotations, annotations were added to the network using the "Annotate nodes" function. Note that these annotations can also be extended using the Perseus annotation files shipped with the software and also available at http://annotations.perseus-framework.org/ within the "PerseusAnnotation" folder. For the purpose of this study we annotate the mouse metabolic network with q-values from each of the omics dataset for which the periodicity analysis was performed using Perseus as explained in the previous sections. The matrix from each periodicity analysis performed on the various omics datasets was exported in .txt format and imported into the Perseus session containing the metabolic network collection using the "Generic matrix upload" function. Since these .txt files are generated using Perseus, upon upload to another Perseus session, Perseus assigns the correct data types to each column in the file automatically. In order to map the q-values from the metabolomics data to the network the ChEBI IDs were used but since there were formatting differences between the network IDs and the data, the "Process text column" function was used to add the string "CHEBI:" to the beginning of the IDs within the metabolomics data with the regular expression "^([^;]+)" and the replacement string "CHEBI:$&" on the "ChEBI" column of the matrix. The column containg the q-values was also renamed using the "Rename columns" function to

"Metabolomics q-value" prior to using the "Annotate nodes" function for mapping the data. The same strategy is used to map the q-values from the phosphoproteomics, proteomics and transcriptomics data. After annotating the network with the q-values, the network was filtered using the "Filter for metabolic reactions" function with the "Number of columns" set to four, "x" set to the metabolomics q-values, "y" set to the phosphoproteomics q-values, "z" set to the proteomics q-values and "a" can be set to the transcriptomics q-values. Subsequently, the four relations are set as "x<0.05", "y<0.1" and "z>=0.33" and no restriction on a. The "Combine through" option is set to "union", since each node type is annotated with a separate column containing the q-values of the cycling analysis for each of the relevant datasets. This results in reactions where either of the three cases is true (cycling metabolite/s, cycling phosphosite/s and non-cycling protein/s). The filtering reduced the network from 6,453 nodes and 48,625 edges to 1,898 nodes and 5,636 edges. For further analysis of the remaining reactions, the "Metabolic reactions to matrix" function was used to collapse each reaction with the filtered nodes to the Perseus matrix format where each row represents a reaction. For this purpose, the column containing the node types were selected, also, the reaction, modifier (enzyme/protein), substrate and products were selected. The resulting matrix is then further processed to filter for reactions that have both phosphosite information for their modifiers and metabolite information for their reactants to reach phosphosite and metabolite pairs of interest (see the Perseus session file named "Analysis-Time-Series.sps).

Network export to third-party software

Any metabolic network analyzed within Perseus can be exported in the .sif (simple interaction file) format using the "SIF export for metabolic reactions" function. For this purpose, both the Perseus metabolic network matrix and network types can be used. For exporting networks from the matrix format, the columns containing the reaction, modifier (enzyme/protein), substrate and products need to be selected along with the path to a Python installation. Exporting the metabolic network from Perseus network format, simply use the sif export function in the network tab (Supplementary Figure 3). The resulting matrix can then be exported and used within third-party software, e.g. Cytoscape.

Supplemental information

Supplementary Table 1.xlsx. Metis analysis of time series multi-omics data summary. The Excel file includes all 111 reactions that have been filtered from the initial mouse metabolic network after mapping all the omics data.

Supplementary Figures.pdf. Supplementary figures 1-3.

Supplementary Notes.pdf. A step by step protocol for using Perseus Metis.

Mendeley Data:
Hamzeiy, Hamid; Ferretti, Daniela; Robles, Maria; Cox, Juergen (2021), "Perseus Metis Data", Mendeley Data, V1, doi: 10.17632/n4nx6x999v.1
http://dx.doi.org/10.17632/n4nx6x999v.1
Metabolic Network Analysis Folder: Perseus session files and data used for Metis analysis for both time series data and static data, along with the final results presented.
Metabolic Networks Folder: Metabolic networks for 11 most common organisms.
Periodicity Analysis Folder: Perseus session files and data used to perform the periodicity analysis on the various datasets.
PHOTON Analysis Folder: Protein-protein interaction networks and Perseus session files used to perform PHOTON analysis for kinases activity prediction.

References

Adamovich, Y., Rousso-Noori, L., Zwighaft, Z., Neufeld-Cohen, A., Golik, M., Kraut-Cohen, J., Wang, M., Han, X., and Asher, G. (2014). Circadian clocks and feeding time regulate the oscillations and levels of hepatic triglycerides. Cell Metab. *19*, 319–330.

Boon, R., Silveira, G.G., and Mostoslavsky, R. (2020). Nuclear metabolism and the regulation of the epigenome. Nat. Metab.

Brown, S.A. (2016). Circadian Metabolism: From Mechanisms to Metabolomics and Medicine. Trends Endocrinol. Metab.

Brüning, F., Noya, S.B., Bange, T., Koutsouli, S., Rudolph, J.D., Tyagarajan, S.K., Cox, J., Mann, M., Brown, S.A., and Robles, M.S. (2019). Sleep-wake cycles drive daily dynamics of synaptic phosphorylation. Science (80-.).

Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. Nat. Rev. Genet.

Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, R., Mittag, F., Schubert, M., Glont, M., Golebiewski, M., et al. (2013). Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Syst. Biol. *7*, 116.

Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M., and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. Nature.

Cox, J., and Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics *13*, S12.

Finger, A.M., Dibner, C., and Kramer, A. (2020). Coupled network of the circadian clocks: a driving force of rhythmic physiology. FEBS Lett.

Fustin, J.M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., Isagawa, T., Morioka, M.S., Kakeya, H., Manabe, I., et al. (2013). XRNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell.

Fustin, J.M., Ye, S., Rakers, C., Kaneko, K., Fukumoto, K., Yamano, M., Versteven, M., Grünewald, E., Cargill, S.J., Tamai, T.K., et al. (2020). Methylation deficiency disrupts biological rhythms from bacteria to humans. Commun. Biol.

Greco, C.M., Cervantes, M., and Fustin, J.-M. (2020). S-adenosyl-l-homocysteine hydrolase links methionine metabolism to the circadian clock and chromatin remodeling. Sci Adv *6*.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. Mol. Cell. Biol.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520.

Hughes, M.E., DiTacchio, L., Hayes, K.R., Vollmers, C., Pulivarthy, S., Baggs, J.E., Panda, S., and Hogenesch, J.B. (2009). Harmonics of circadian gene transcription in mammals. PLoS Genet. *5*, e1000442.

Krishnaiah, S.Y., Wu, G., Altman, B.J., Growe, J., Rhoades, S.D., Coldren, F., Venkataraman, A., Olarerin-George, A.O., Francey, L.J., Mukherjee, S., et al. (2017). Clock Regulation of Metabolites Reveals Coupling between Transcription and Metabolism. Cell Metab. *25*, 961-974.e4.

Li, X., Egervari, G., Wang, Y., Berger, S.L., and Lu, Z. (2018). Regulation of chromatin and gene expression by metabolic enzymes and metabolites. Nat. Rev. Mol. Cell Biol.

Mews, P., Donahue, G., Drake, A.M., Luczak, V., Abel, T., and Berger, S.L. (2017). Acetyl-CoA synthetase regulates histone acetylation and hippocampal memory. Nature.

Morloy, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S., and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature.

Needham, E.J., Parker, B.L., Burykin, T., James, D.E., and Humphrey, S.J. (2019). Illuminating the dark phosphoproteome. Sci. Signal.

Neufeld-Cohen, A., Robles, M.S., Aviram, R., Manella, G., Adamovich, Y., Ladeuix, B., Nir, D., Rousso-Noori, L., Kuperman, Y., Golik, M., et al. (2016). Circadian control of oscillations in mitochondrial rate-limiting enzymes and nutrient utilization by PERIOD proteins. Proc. Natl. Acad. Sci. U. S. A.

Panda, S., Antoch, M.P., Miller, B.H., Su, A.I., Schook, A.B., Straume, M., Schultz, P.G., Kay, S.A., Takahashi, J.S., and Hogenesch, J.B. (2002). Coordinated transcription of key pathways in the mouse by the circadian clock. Cell.

Pandhare, J., Donald, S.P., Cooper, S.K., and Phang, J.M. (2009). Regulation and function of proline oxidase under nutrient stress. J. Cell. Biochem.

Robitaille, A.M., Christen, S., Shimobayashi, M., Cornu, M., Fava, L.L., Moes, S., Prescianotto-Baschong, C., Sauer, U., Jenoe, P., and Hall, M.N. (2013). Quantitative phosphoproteomics reveal mTORC1 activates de novo pyrimidine synthesis. Science (80-.). *339*, 1320–1323.

Robles, M.S., Cox, J., and Mann, M. (2014). In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. PLoS Genet. *10*, e1004047.

Robles, M.S., Humphrey, S.J., and Mann, M. (2017). Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology. Cell Metab. *25*, 118–127.

Rudolph, J.D., and Cox, J. (2019). A Network Module for the Perseus Software for Computational Proteomics Facilitates Proteome Interaction Graph Analysis. J. Proteome Res. *18*, 2052–2064.

Rudolph, J.D., de Graauw, M., van de Water, B., Geiger, T., and Sharan, R. (2016). Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. Cell Syst.

Storch, K.F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F.C., Wong, W.H., and Weitz, C.J. (2002). Extensive and divergent circadian gene expression in liver and heart. Nature.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

Ueda, H.R., Chen, W., Adachi, A., Wakamatsu, H., Hayashi, S., Takasugi, T., Nagano, M., Nakahama, K.I., Suzuki, Y., Sugano, S., et al. (2002). A transcription factor response element for gene expression during circadian night. Nature.

Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D.P., Zecha, J., Asplund, A., Li, L., Meng, C., et al. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol. Syst. Biol.

Yu, S.H., Ferretti, D., Schessner, J.P., Rudolph, J.D., Borner, G.H.H., and Cox, J. (2020). Expanding the Perseus Software for Omics Data Analysis With Custom Plugins. Curr.

Protoc. Bioinforma.

Figure legends

Figure 1. Perseus framework and Metis toolbox. Perseus is a plugin-based software for the analysis of omics data. It supports two basic data types, matrices and networks. The former usually carries relative bimolecular concentration data and a multitude of activities exist in Perseus to process them. The generic network data structure consists of annotated nodes and edges to accommodate diverse biological networks. Standard plugins contain activities for the creation, import processing and analysis of, for instance, protein-protein interaction networks or networks consisting of kinase-substrate relations. The Metis toolbox extends the Perseus network framework to metabolic pathways consisting of reactions, connecting metabolites that are consumed or created with the catalyzing enzymes. Annotation of these networks connects metabolomics matrix data with matrices related to enzymes, which can be proteomics, transcriptomics or phosphoproteomics data. Network nodes and edges can then be filtered with simultaneous criteria on multiple omics types resulting in a metabolic reaction matrix containing the results of interest. Finally, results can be exported in .sif format, for instance for visualization in Cytoscape.

Figure 2. Schematic overview of the data analysis workflow with time-series circadian data. Datasets used from five different studies of circadian rhythms in mice liver serve as input, along with the results of the kinase activity prediction using the PHOTON plugin within the Perseus software package. These were independently reanalyzed using the periodicity analysis toolkit in the Perseus software. The results were then merged with the reconstructed mouse metabolic network from the BioModels database using the network analysis module of Perseus.

Fig. 3. Results of cycling analysis for the individual omics datasets. a. Pie charts show the percentage of cycling transcripts, proteins, predicted kinases, phosphosites, metabolites and lipids with respect to the total dataset. **b.** Heat maps of cycling molecules are shown for each omics type. Biomolecules are sorted vertically according to their circadian phase while horizontally the time points were mapped and averaged to a 24h interval in case the time series were longer. **c.** Rose plots indicating peaking positions of all cycling molecules in a circular histogram.

Fig. 4. Representation and processing of metabolic networks within Metis. a. Exemplary network structure which is represented in Metis using the three node types reaction, metabolite and enzyme, which are connected according to the participation of biomolecules in reactions. **b.** Nodes are attached to annotations originating from matrices filled with quantitative omics data. The metabolite nodes are affiliated with rows in a matrix object carrying metabolite concentrations over potentially complex experimental designs, which are time series in the case at hand. The enzyme nodes are primarily associated with proteomics data, connecting nodes to the quantitative data on relative enzyme concentrations, here also in the form of time series data. Also quantitative posttranslational modification data, as for instance phosphorylation or mRNA level data are mapped here to the enzyme nodes. After mapping various omics data to the network, filters can be applied on each node (filtered nodes are shown in green in this example). **c.** Applying the filter results in a network where only the relevant reactions and enzymes are retained.

Fig. 5. Network of enzymatic reactions with oscillating enzyme phosphorylation, substrates and products. Orange squares depict enzymes and green hexagons depict metabolites. For the highlighted regions common profile plots of normalized phosphosite and metabolite intensities are shown. **a.** ACOX1 and PRODH. **b.** UBP1. **c.** HARS2. **d.** GNMT. **e.** PTDSS1. **f.** ACSS2, PLA2G4A and COASY.

Fig. 6. Histogram of the difference between acrophases of metabolites and phosphosites. All pairs of cycling metabolites and phosphorylation sites resulting from the network filtering, that are hence connected through a reaction were used to create a histogram of phase differences. The differences between acrophases of metabolites and corresponding enzyme phosphorylations are sorted into 3h bins. Enrichments can be observed around the ‘in phase’ and ‘in anti-phase’ regions.

Table legends

Table 1. Datasets used in this study. Details of the five omics datasets of circadian mouse liver entrained in day-night cycles and then free running from time point 0. Several details of the time series acquisition, such as the total acquisition time, the sampling rate and the number of replicates per time point vary. Cycling q-values differed between the analyses performed in the respective publications. In order to keep consistency with previous work, we applied the cycling q-value that was used in each publication.

Figure 1

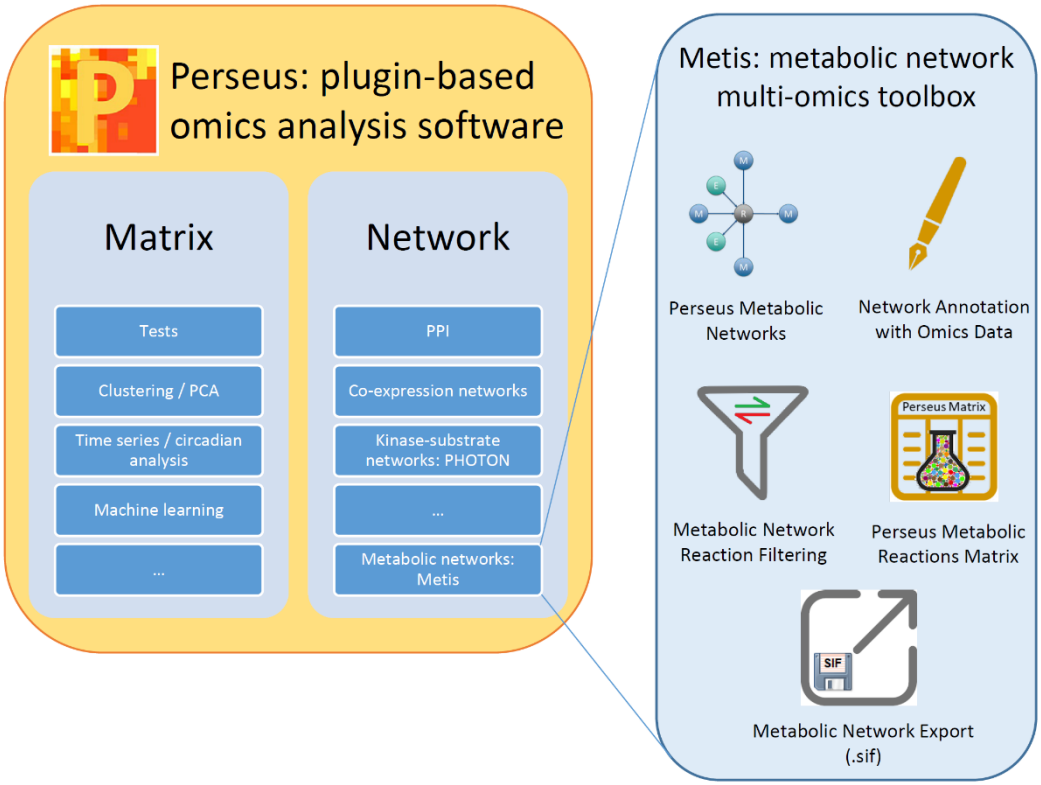

Perseus: plugin-based omics analysis software
Matrix
Tests
Clustering / PCA
Time series / circadian analysis
Machine learning
...
Network
PPI
Co-expression networks
Kinase-substrate networks: PHOTON
...
Metabolic networks: Metis
Metis: metabolic network multi-omics toolbox
Perseus Metabolic Networks
Network Annotation with Omics Data
Metabolic Network Reaction Filtering
Perseus Matrix
Perseus Metabolic Reactions Matrix
SIF
Metabolic Network Export (.sif)

Figure 2

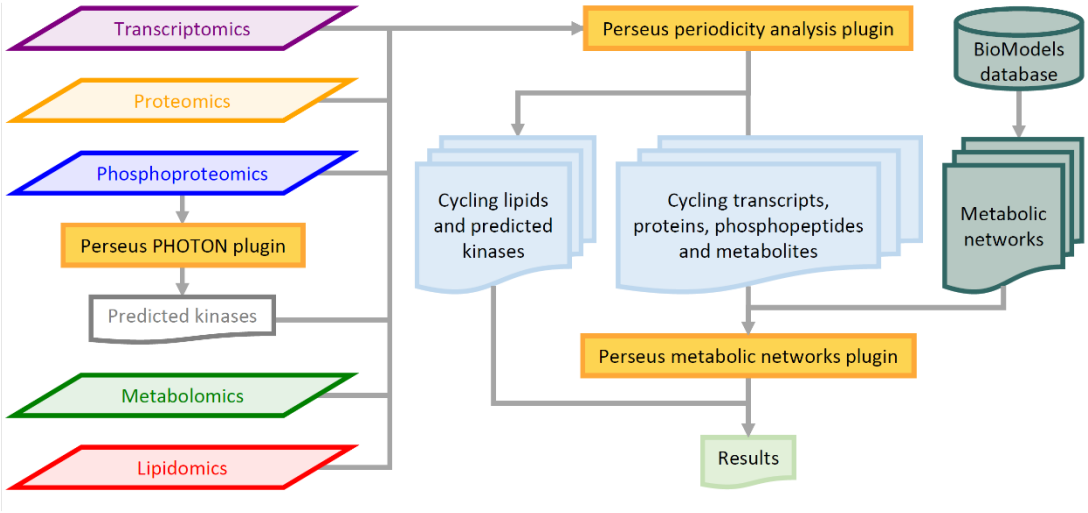

Transcriptomics
Proteomics
Phosphoproteomics
Perseus PHOTON plugin
Predicted kinases
Metabolomics
Lipidomics
Perseus periodicity analysis plugin
BioModels database
Cycling lipids and predicted kinases
Cycling transcripts, proteins, phosphopeptides and metabolites
Metabolic networks
Perseus metabolic networks plugin
Results

Figure 3

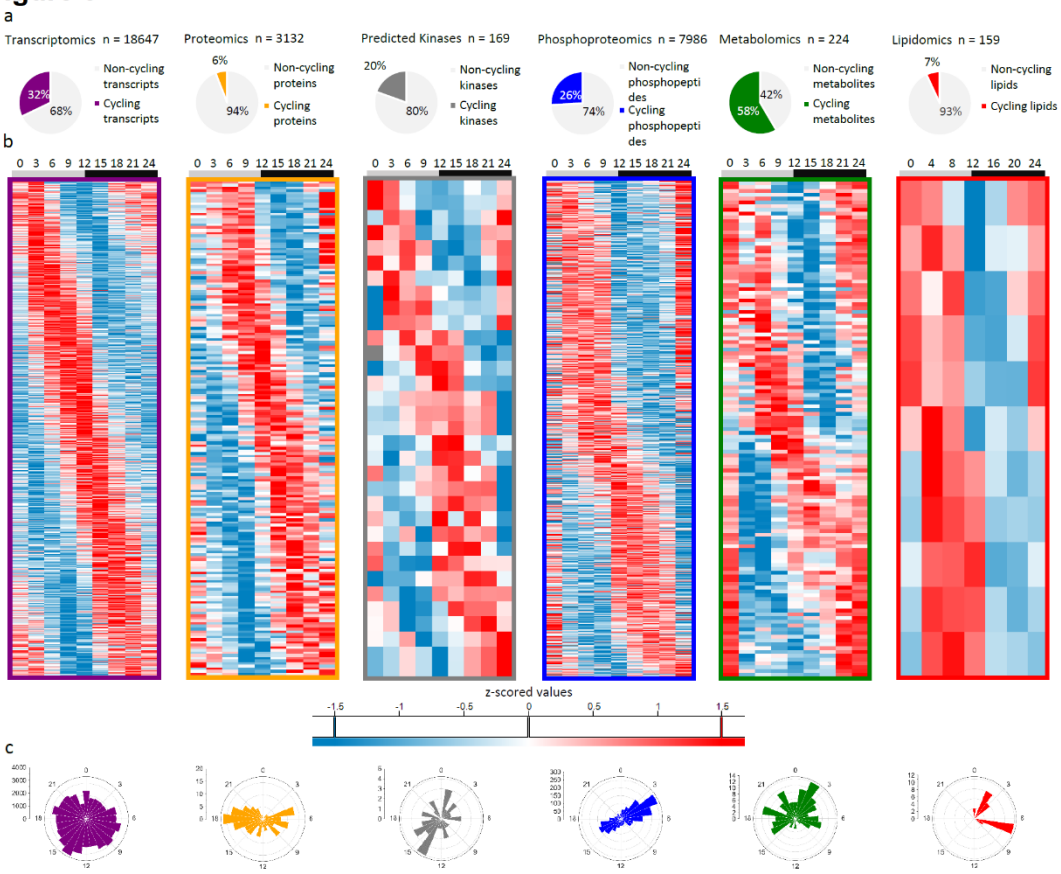

a
Transcriptomics n = 18647
32%
68%
Non-cycling transcripts
Cycling transcripts
Proteomics n = 3132
6%
94%
Non-cycling proteins
Cycling proteins
Predicted Kinases n = 169
20%
80%
Non-cycling kinases
Cycling kinases
Phosphoproteomics n = 7986
26%
74%
Non-cycling phosphopeptides
Cycling phosphopeptides
Metabolomics n = 224
42%
58%
Non-cycling metabolites
Cycling metabolites
Lipidomics n = 159
7%
93%
Non-cycling lipids
Cycling lipids
b
0 3 6 9 12 15 18 21 24
0 4 8 12 16 20 24
z-scored values
-1.5 -1 -0.5 0 0.5 1 1.5
c

Figure 4

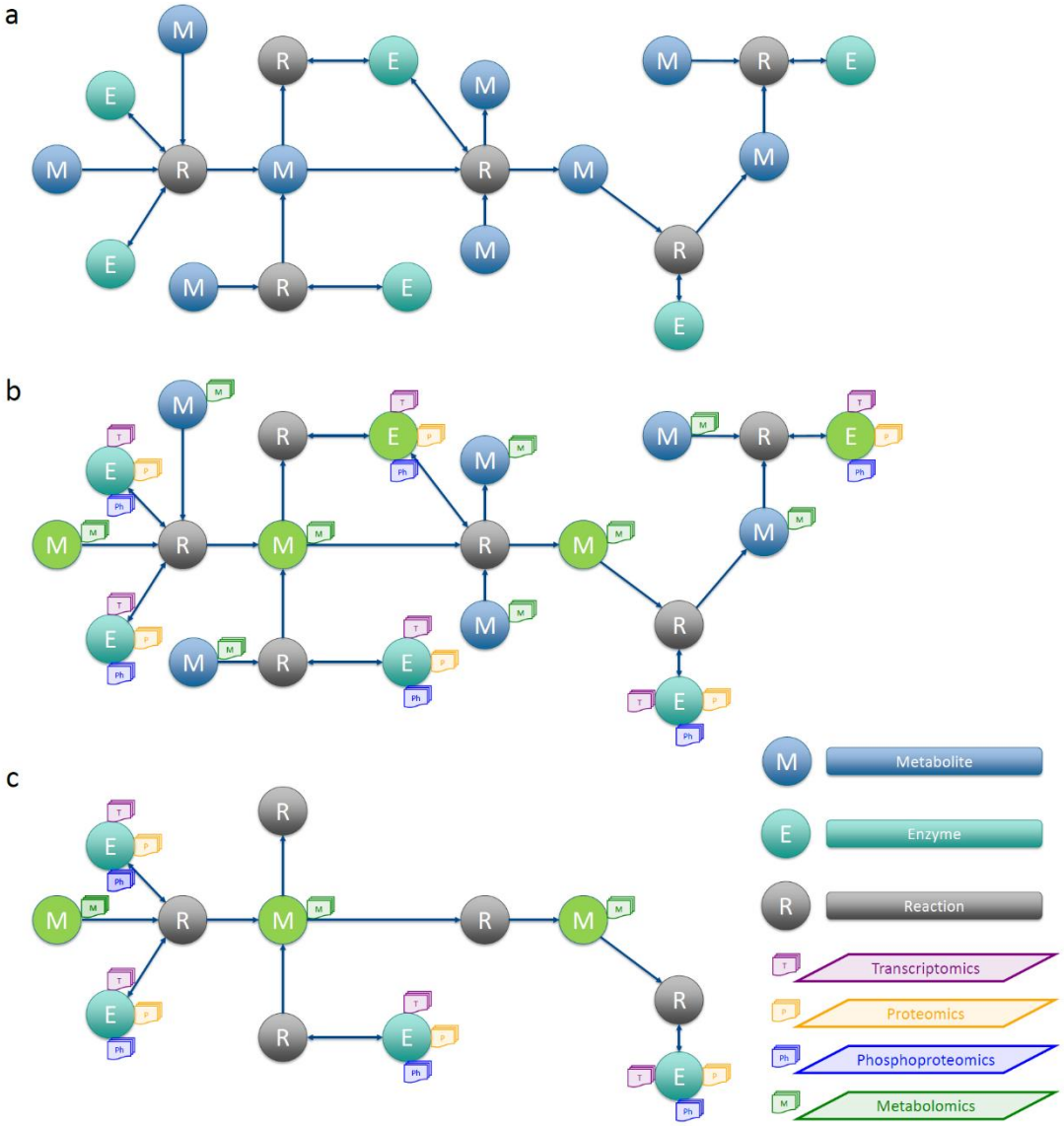
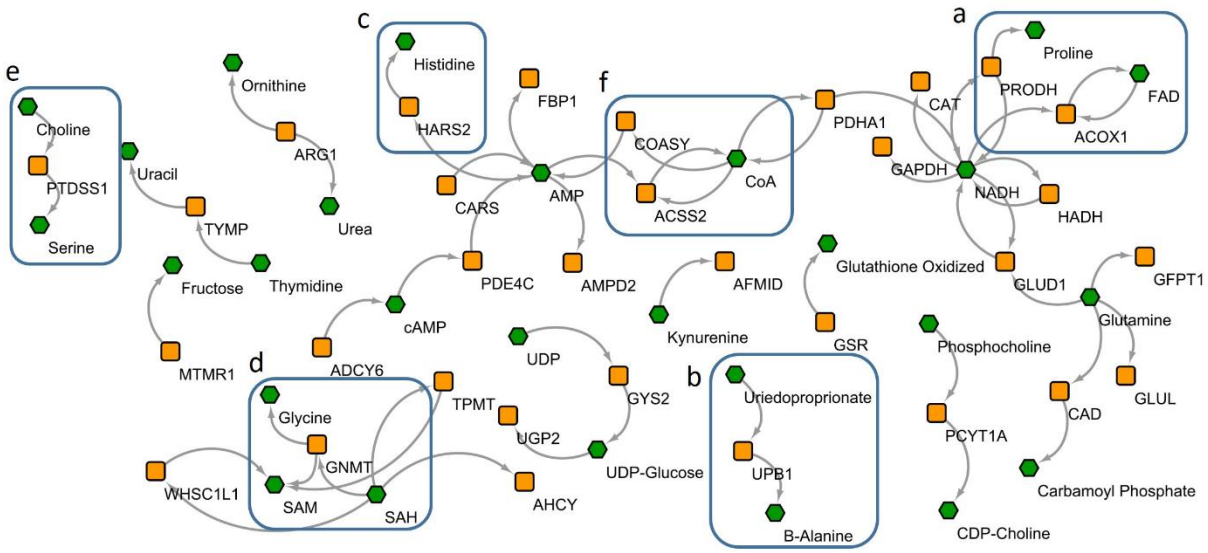

a
b
c
Metabolite
Enzyme
Reaction
Transcriptomics
Proteomics
Phosphoproteomics
Metabolomics

Figure 5


a
Proline
PRODH
FAD
ACOX1
b
Uriedoproprionate
UPB1
B-Alanine
c
Histidine
HARS2
d
Glycine
GNMT
SAM
SAH
e
Choline
PTDSS1
Serine
f
COASY
ACSS2
CoA
Ornithine
ARG1
Urea
Uracil
TYMP
Thymidine
Fructose
MTMR1
FBP1
CARS
AMP
PDE4C
AMPD2
cAMP
ADCY6
TPMT
UDP
GYS2
UGP2
UDP-Glucose
AHCY
WHSC1L1
PDHA1
CAT
GAPDH
NADH
HADH
Glutathione Oxidized
GSR
AFMID
Kynurenine
GLUD1
GFPT1
Glutamine
Phosphocholine
PCYT1A
CDP-Choline
CAD
GLUL
Carbamoyl Phosphate

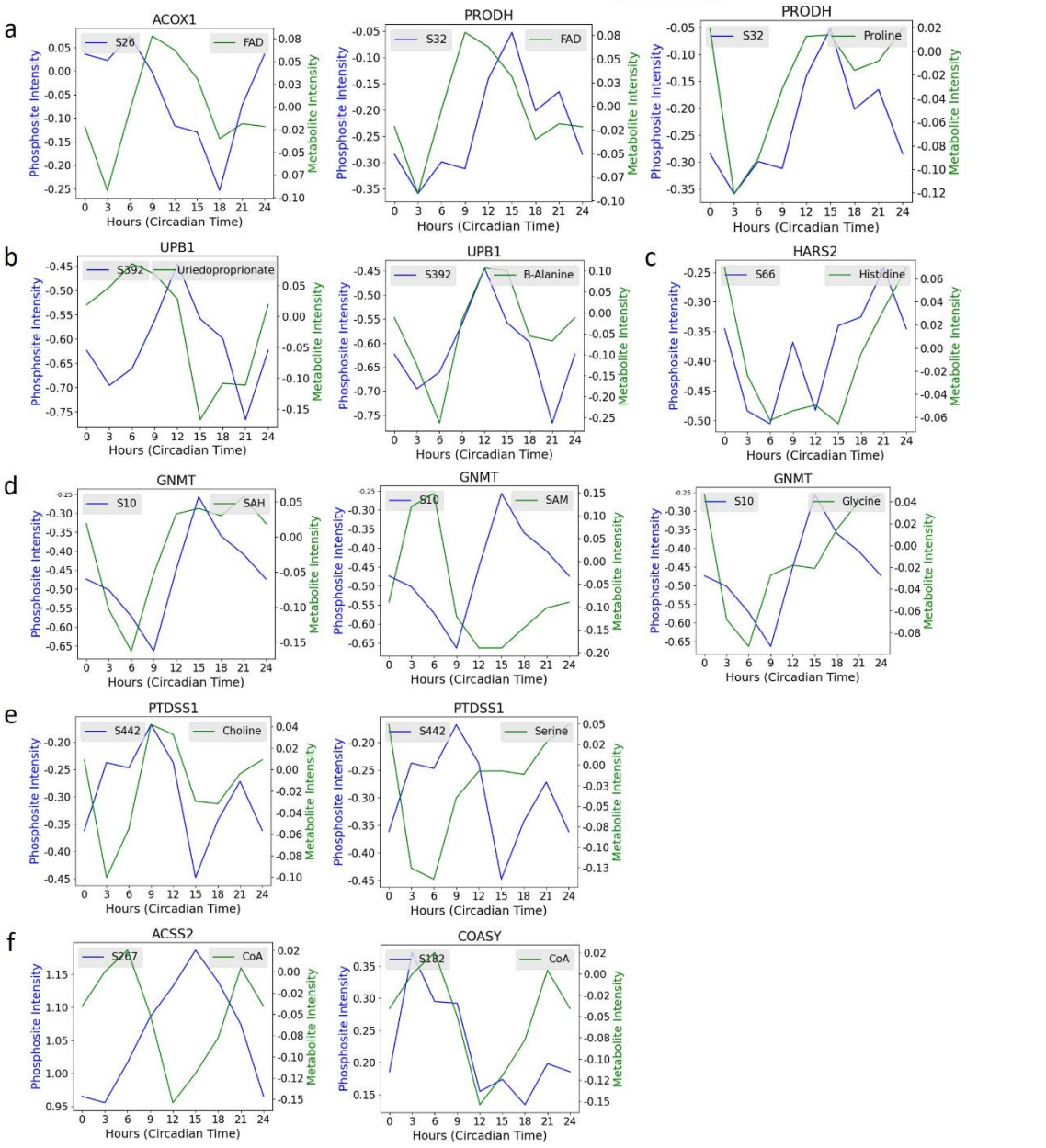

a
ACOX1
S26
FAD
PRODH
S32
FAD
PRODH
S32
Proline
b
UPB1
S392
Uriedoproprionate
UPB1
S392
B-Alanine
c
HARS2
S66
Histidine
d
GNMT
S10
SAH
GNMT
S10
SAM
GNMT
S10
Glycine
e
PTDSS1
S442
Choline
PTDSS1
S442
Serine
f
ACSS2
S267
CoA
COASY
S182
CoA
Phosphosite Intensity
Metabolite Intensity
Hours (Circadian Time)

Figure 6

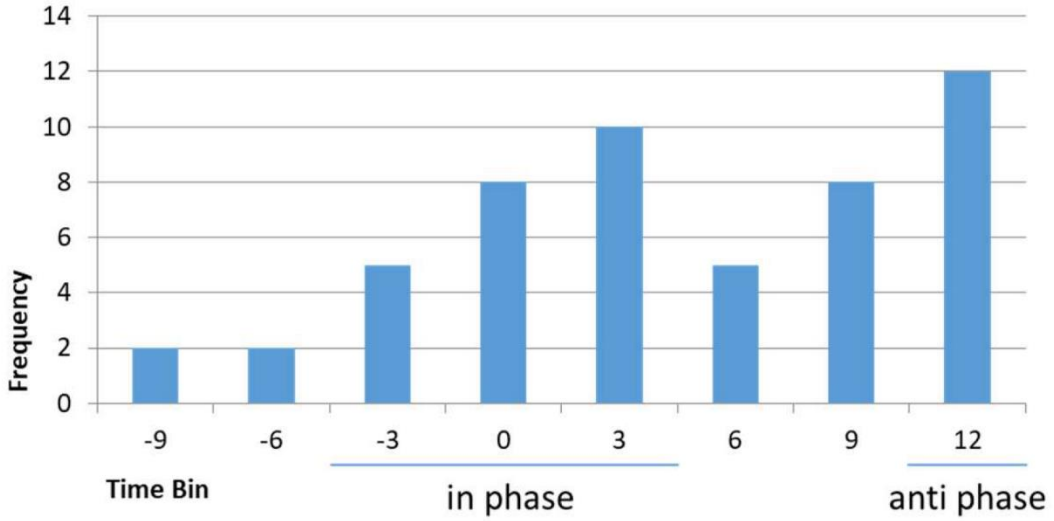
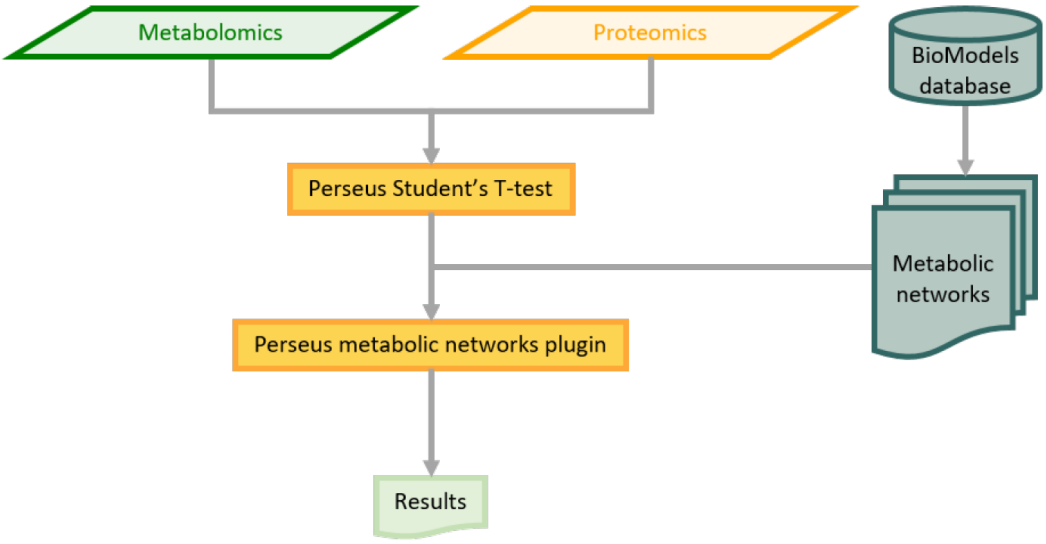
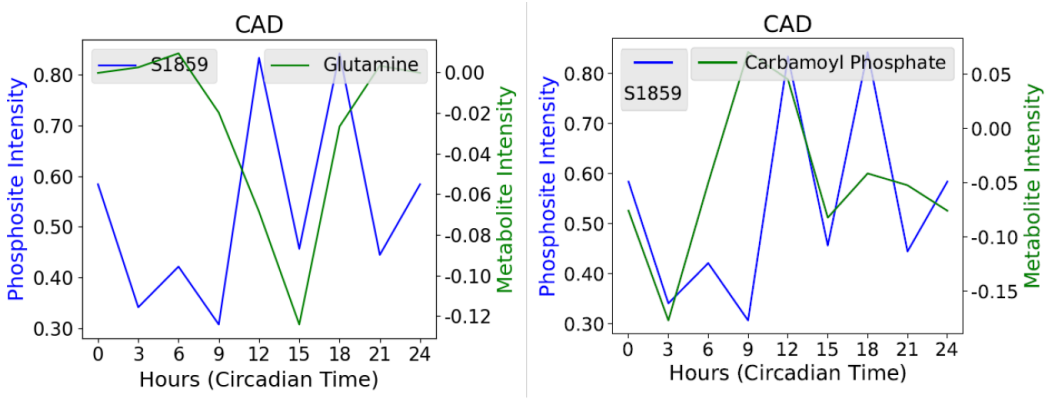

14
12
10
8
6
4
2
0
Frequency
-9
-6
-3
0
3
6
9
12
Time Bin
in phase
anti phase

Table 1. Datasets used in this study. Details of the five omics datasets of circadian mouse liver entrained in day-night cycles and then free running from time point 0. Several details of the time series acquisition, such as the total acquisition time, the sampling rate and the number of replicates per time point vary. Cycling q-values differed between the analyses performed in the respective publications. In order to keep consistency with previous work, we applied the cycling q-value that was used in each publication.

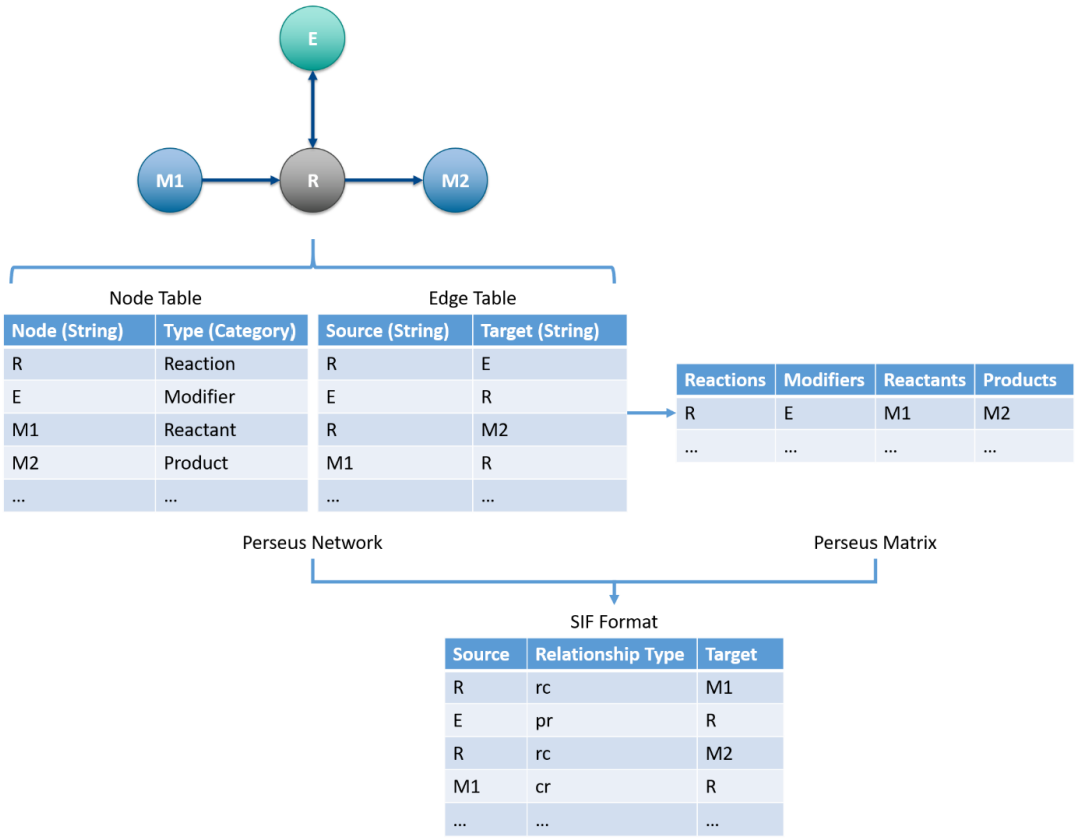
Paper	Hughes, M. E. et al. 2009	Robles, M. S., Cox, J. & Mann, M. 2014	Robles, M. S., Humphrey, S. J. & Mann, M. 2017	Krishnaiah, S. Y. et al. 2017	Adamovich, Y. et al. 2014
Dataset	Transcriptomics	Proteomics	Phosphoproteomics	Metabolomics	Lipidomics
Number of identified molecules	18647	3132	7986	224	159
Number of cycling molecules	5989	186	2066	131	11
Total duration of sampling (hours)	48	48	48	48	20
Number of time points measured	48 (every hour)	16 (every three hours)	16 (every three hours)	48 (every hour)	6 (every four hours)
Replicates per time point	1	3	3	4	4
Cycling q-value threshold	0.05	0.33	0.1	0.05	0.05



Supplementary Figure 1: Schematic overview of the data analysis workflow with static data with two conditions. As an example for an analysis done with multi-omics data with two conditions as shown in the figure, the liver metabolomics and proteomics data which were processed as explained in the methods section of the paper prior to periodicity analysis was utilized. The time points corresponding to the circadian times 6 and 18 (including 30 and 42) were chosen and performed a Student's T-test between these times using Perseus's "Two-sample tests" function, with the default values. Later, the mouse metabolic network which was processed and prepared as explained in the methods section was mapped and filtered with the significant metabolites and proteins. This is done by using the "Annotate node" function of Perseus to map the Student's T-test q-values to the network and the "Filter for metabolic reactions" function of Perseus for filtering the network for reactions of interest. The resulting network can then be converted to a Perseus matrix using the "Metabolic reactions to matrix" function of Perseus for further analysis (see Perseus session file named "Analysis-Static.sps").



Supplementary Figure 2: CAD S1859, glutamine and carbamoyl phosphate data averaged over 24 hours. CAD is an enzyme within the pyrimidine pathway with several cofactor binding sites for Zn^{2+} and Mg^{2+}, which is also known to be allosterically regulated by phosphorylation at S1859 (cycling at q-value 0.07 and phase 15.74) by RPS6KB1, which stimulates dihydroorotase activity, downstream of MTOR. Several phosphosites are detected for RPS6KB1 and three cycle at q-values at about 0.002 and phases at about 2.75 (S441, T444 and S447). This is also the case for MTOR, for which S2448 cycles at q-value 0.004 and phase 16.75. Phosphorylation at S1406 for CAD reduces sensitivity to feedback inhibition by UTP, but this phosphosite is not detected within the phosphoproteomics data. CAD's transcript is detected within the transcriptomics data as non-cycling with a q-value of 0.29 but it is not detected within the proteomics data. At the metabolomics level glutamine (q-value 0 and phase 2.73) and carbamoyl phosphate (q-value 0 and phase 12.44) are cycling as reactant and product of the reaction, 1 H2O + 2 ATP + 1 L-glutamine zwitterion + 1 bicarbonate = 2 H(+) + 1 carbamoyl phosphate(2-) + 2 ADP + 1 L-glutamate + 1 phosphate, where CAD has catalytic activity, respectively. The data shows that as phosphorylation levels increase at CAD S1859, glutamine levels decrease and carbamoyl phosphate levels increase. It is interesting that the kinase responsible for CAD S1859, RPS6KB1, has a phase similar to glutamine, and that MTOR S2448 has a phase similar to CAD S1859.

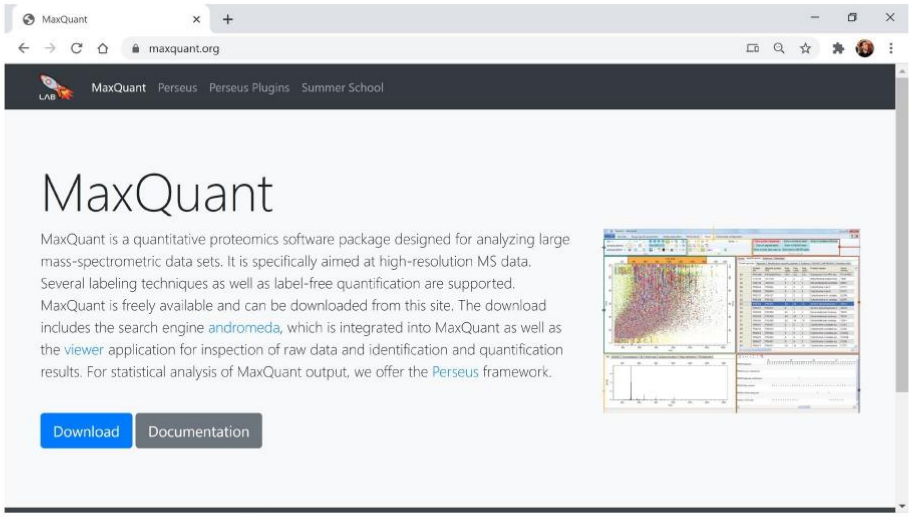


Supplementary Figure 3: Schematic overview of the Perseus "Metabolic reactions to matrix" function and SIF format export functions. The Perseus "Metabolic reactions to matrix" function allows the user to convert the Perseus network format structure to a Perseus matrix, where each reaction within the respective network is represented as a row in the matrix output along with its modifiers, reactants and products. This allows the reactions within the network to be converted to a human readable format, and also allows the user to be able to use the vast array of functions available within Perseus for matrix data manipulation on the reactions of the network. The "Metabolic reactions to matrix" function, takes user input on the relevant reaction, modifier, reactant and product nodes, and works to take all the reaction nodes within the network and annotates them with columns corresponding to their respective modifiers, reactants and products. Perseus metabolic networks and matrices can also be converted to the SIF format and exported to third-party software for further analysis or visualization using the "SIF export for metabolic reactions" function of Perseus.

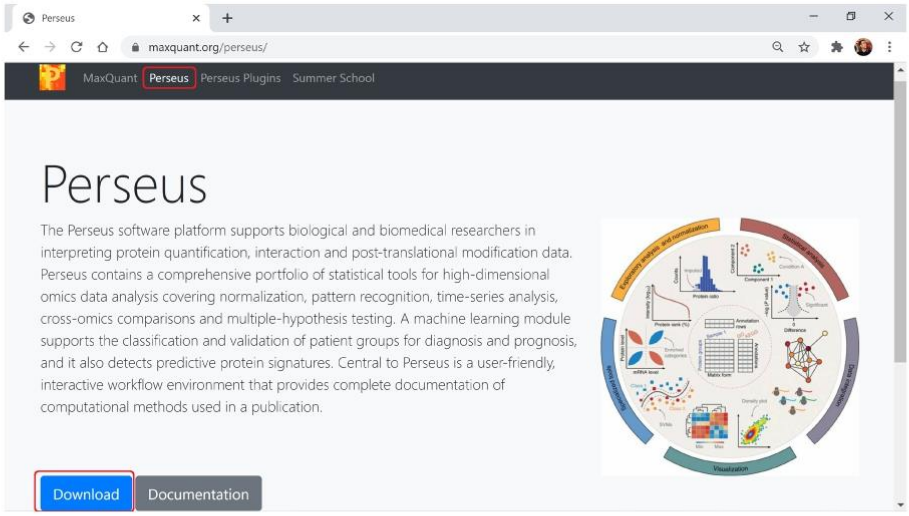
Supplementary Notes

Perseus Multiomics Tutorial

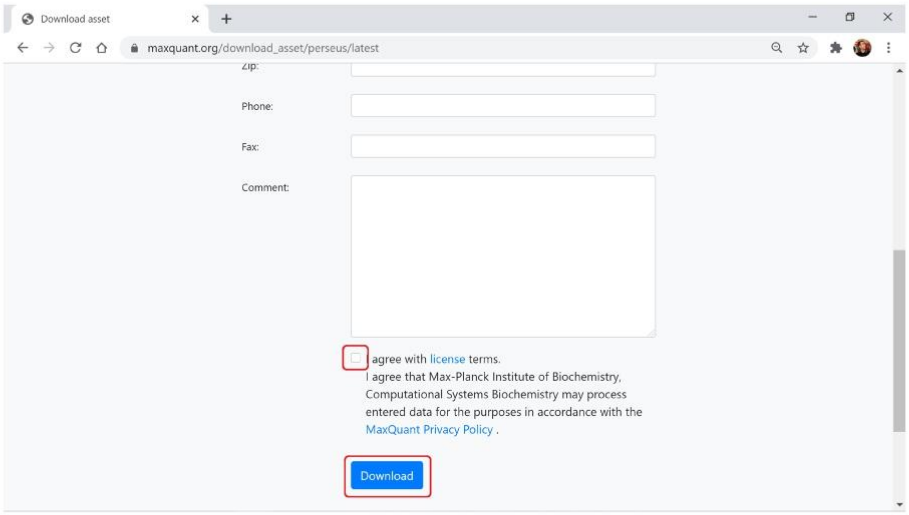
1. Navigate to https://maxquant.org/ to download Perseus:



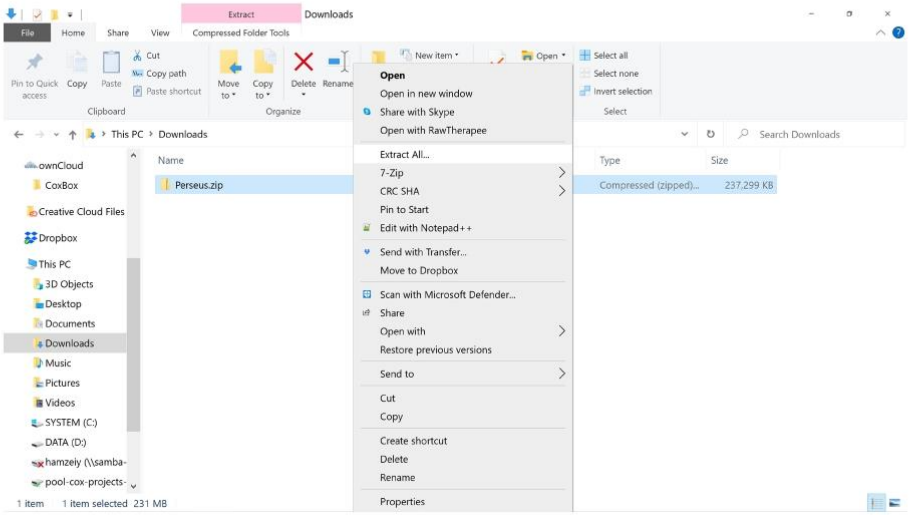
2. Click on “Perseus” on the top menu to navigate to the Perseus page and click on the “Download” button:



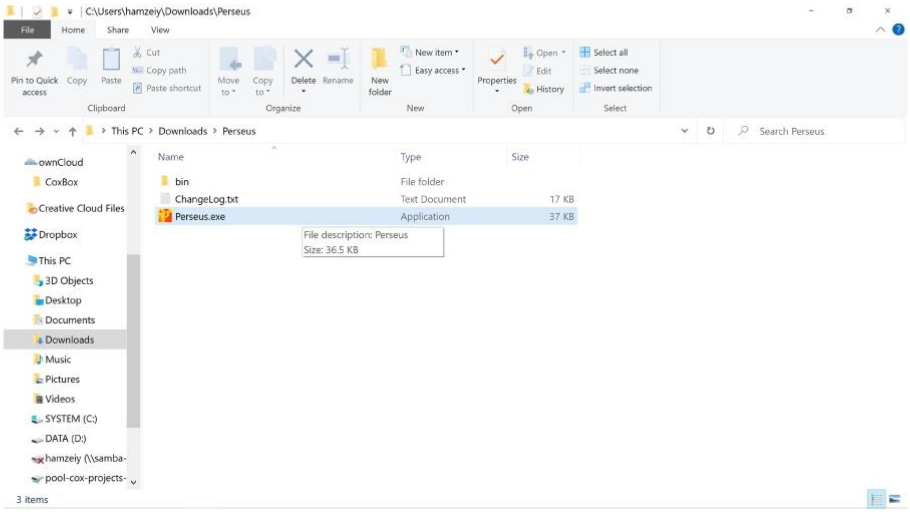
3. Fill in the form, click on the check box to agree with Perseus's license and click on the download button:



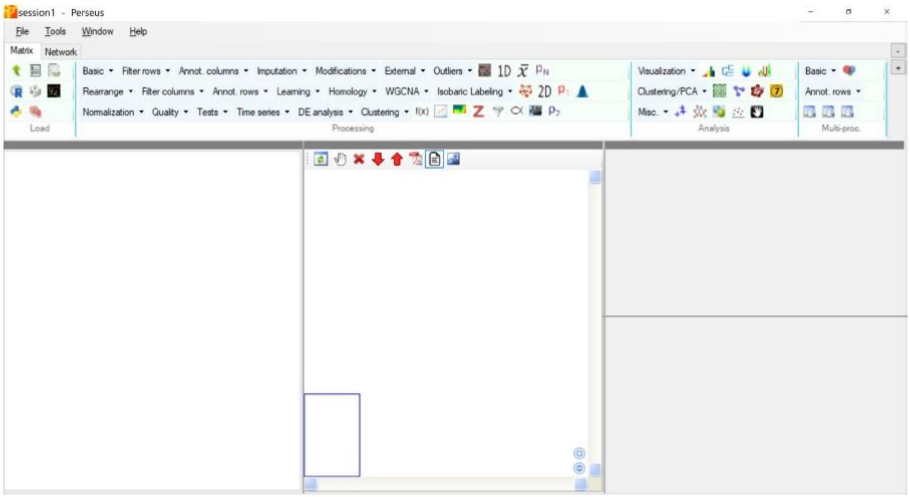
4. Locate the downloaded .zip file and extract its content:



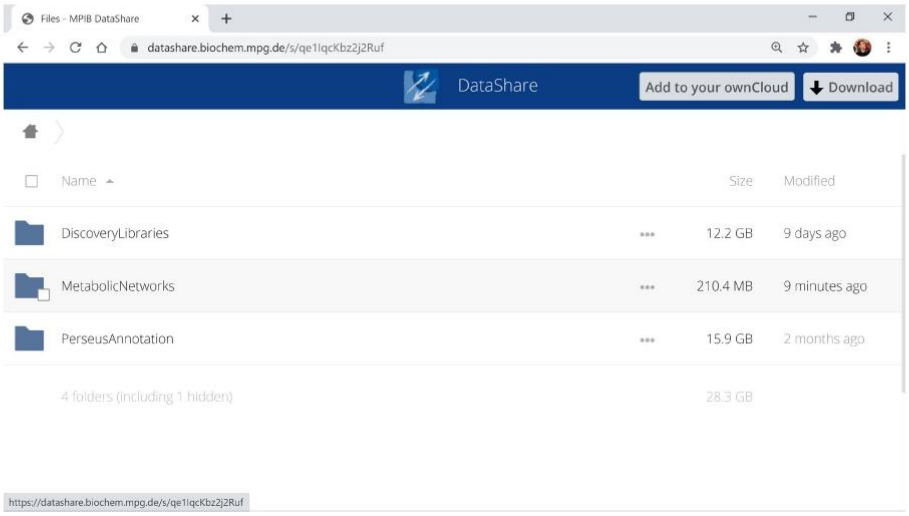
5. Navigate to the extracted folder and double click on "Perseus.exe" to launch Perseus:



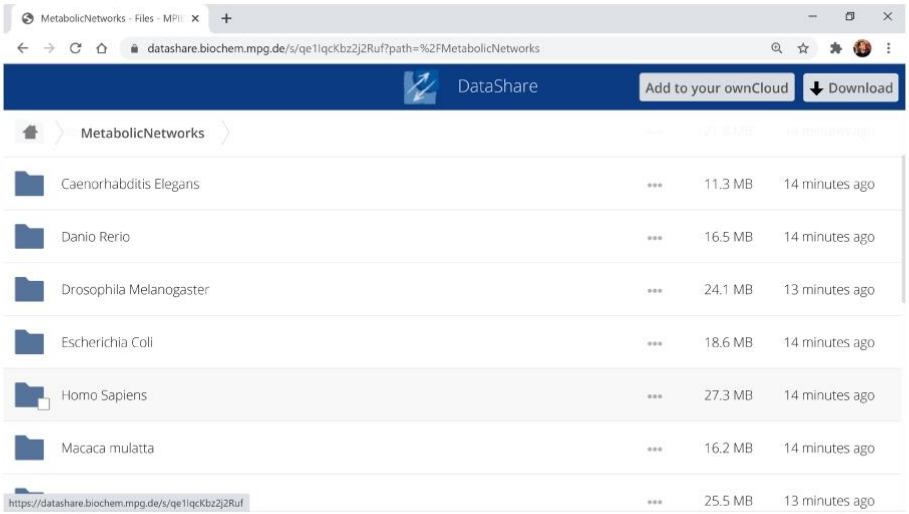
6. You now have Perseus open:



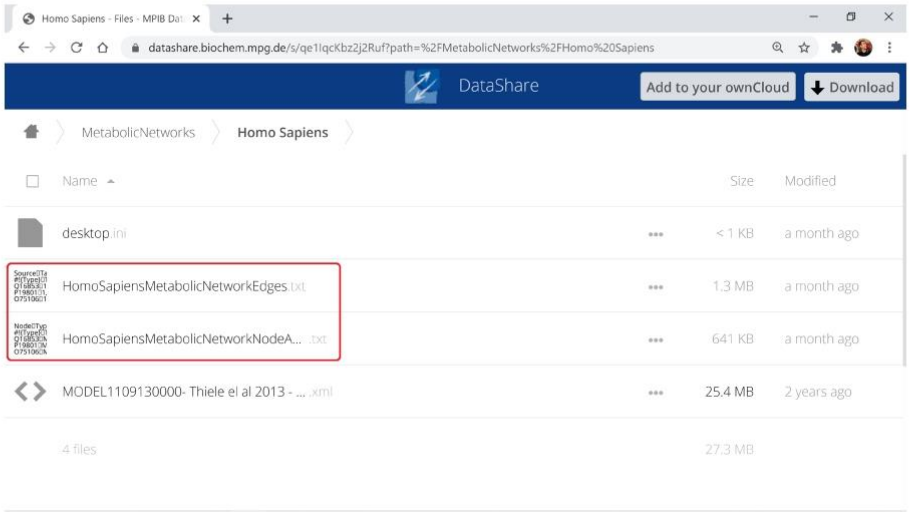
7. Navigate to http://annotations.perseus-framework.org/ to download your metabolic network of choice before uploading it to Perseus for your analysis:



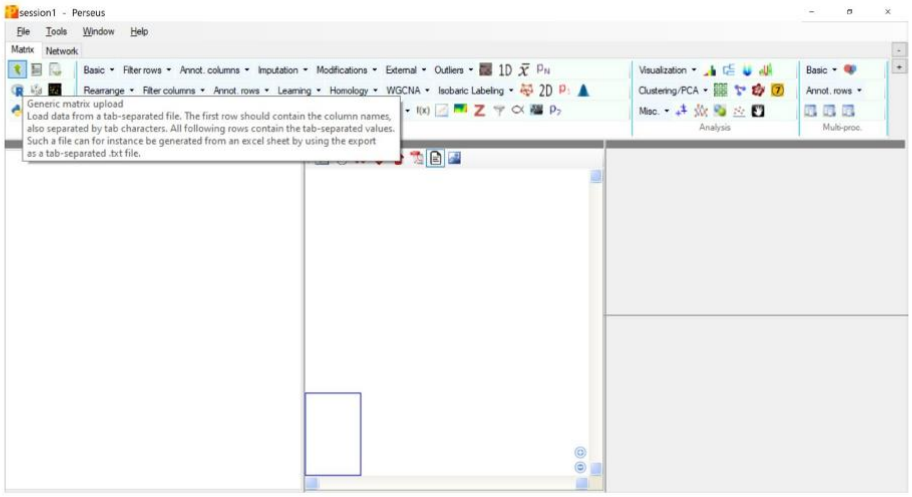
8. Navigate to the "MetabolicNetworks" folder:



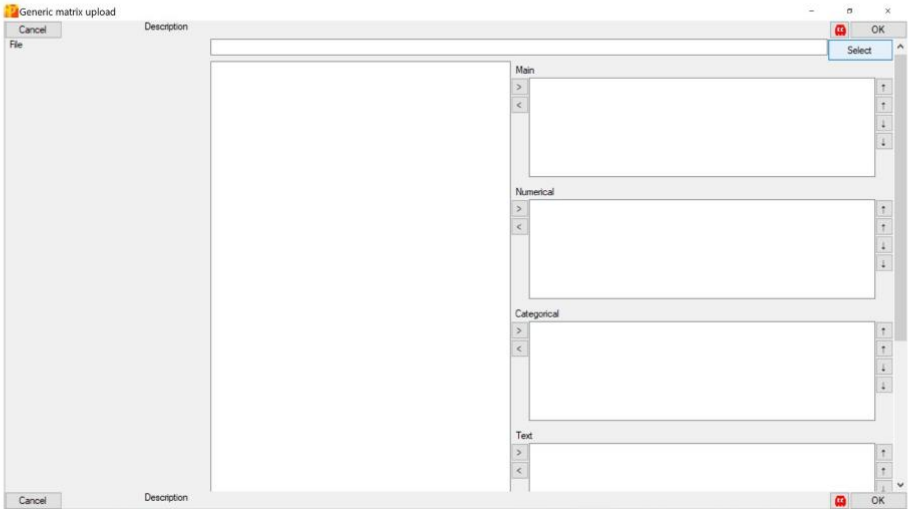
9. Choose and navigate to your organism of choice and download the two .txt files corresponding to the edges of the network and node annotations (if the organism you are interested in is not in the list, contact cox@biochem.mpg.de):



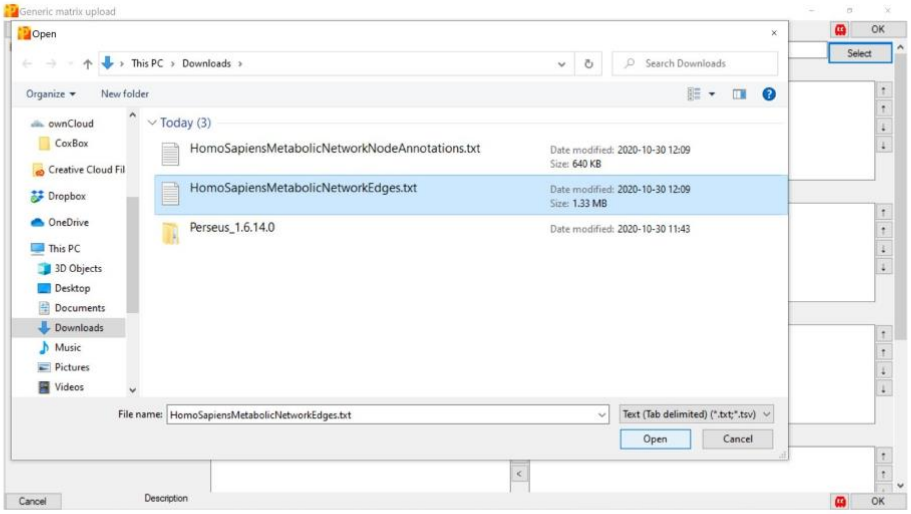
10. After downloading your network of choice, go back to Perseus and click on the "Generic matrix upload" button in Perseus:



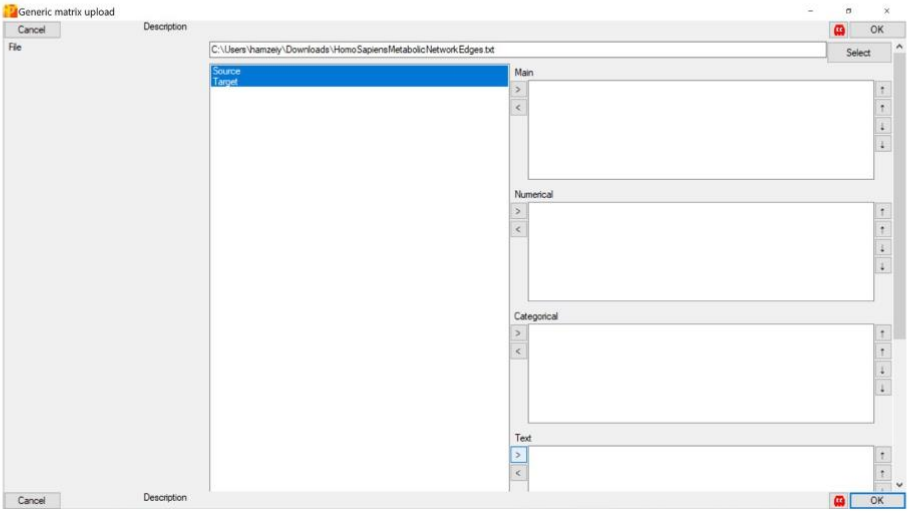
11. On the "Generic matrix upload" dialog box click on "Select" to select and import the network files:



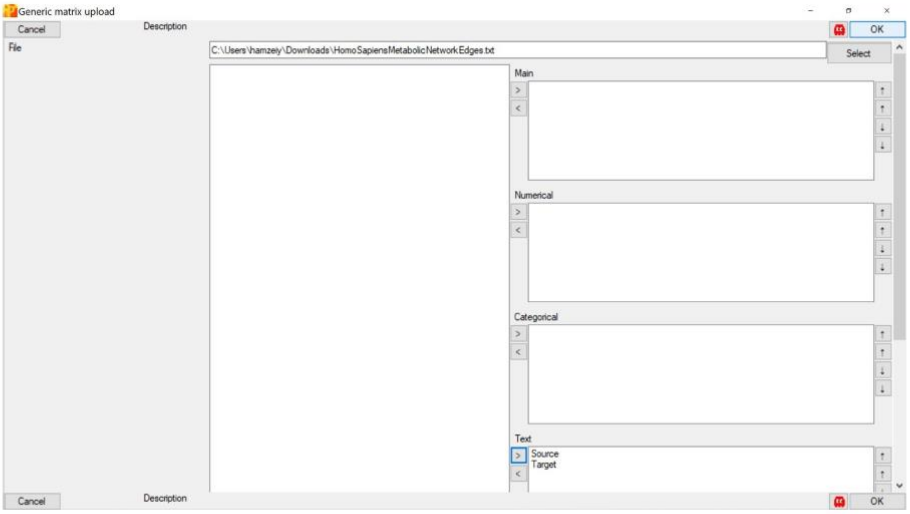
12. First select the .txt file corresponding to the network edges:



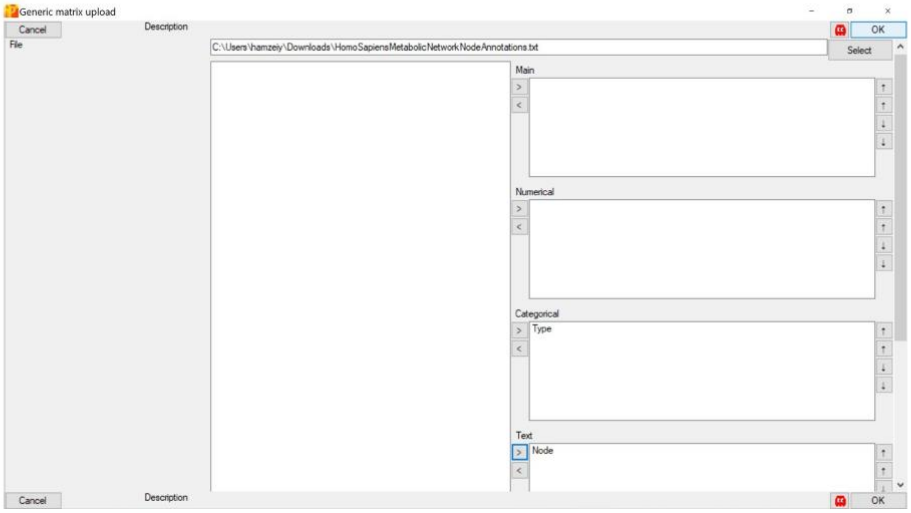
13. Both the “Source” and “Target” columns should be selected and moved to the “Test” box on the right side of the dialog box:



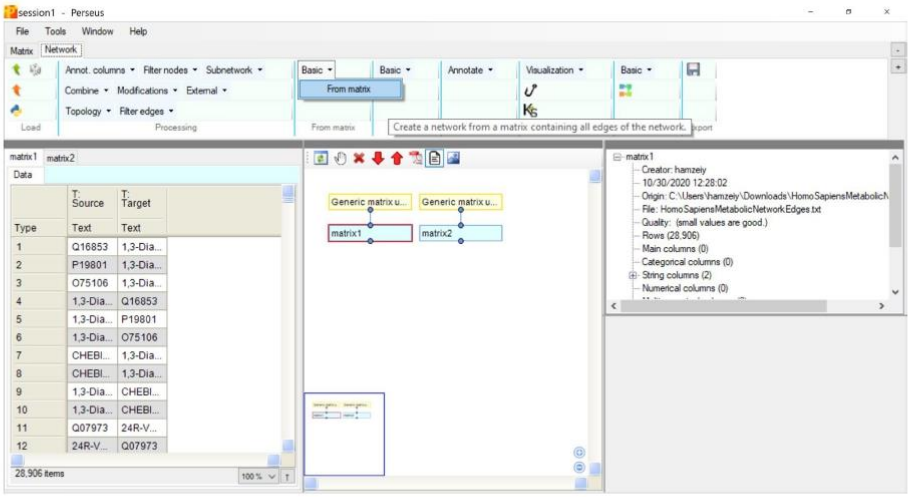
14. After moving the “Source” and “Target” columns to the “Text” box, press “OK”:



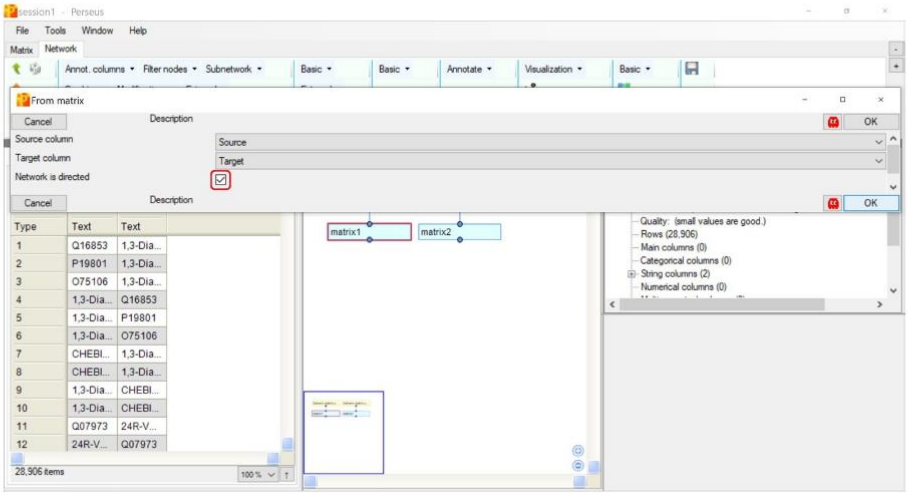
15. Repeat steps 10-14 for the node annotations .txt files, keeping in mind that “Node” should be in the “Text” box and “Type” should be in the “Categorical” box:



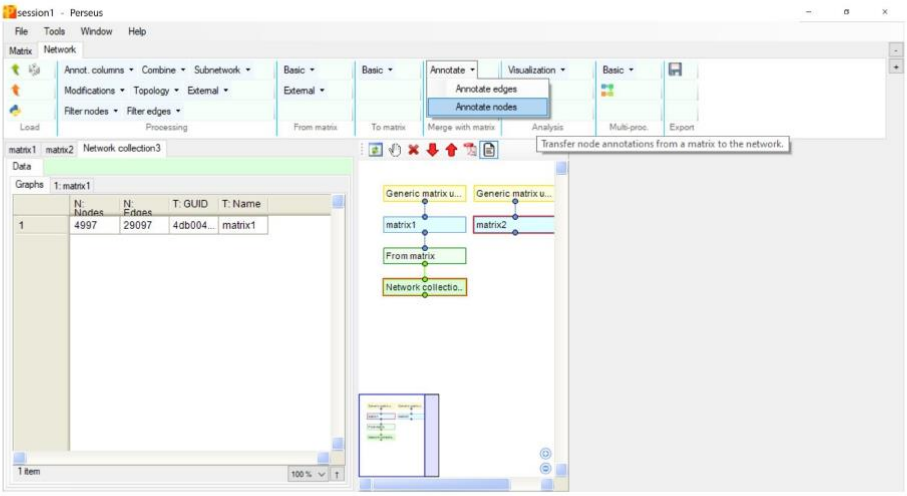
16. Select the matrix with the network edges which you imported at step 14 and from the "Network" tab, within the "From matrix" section, select "Basic" and "From matrix":



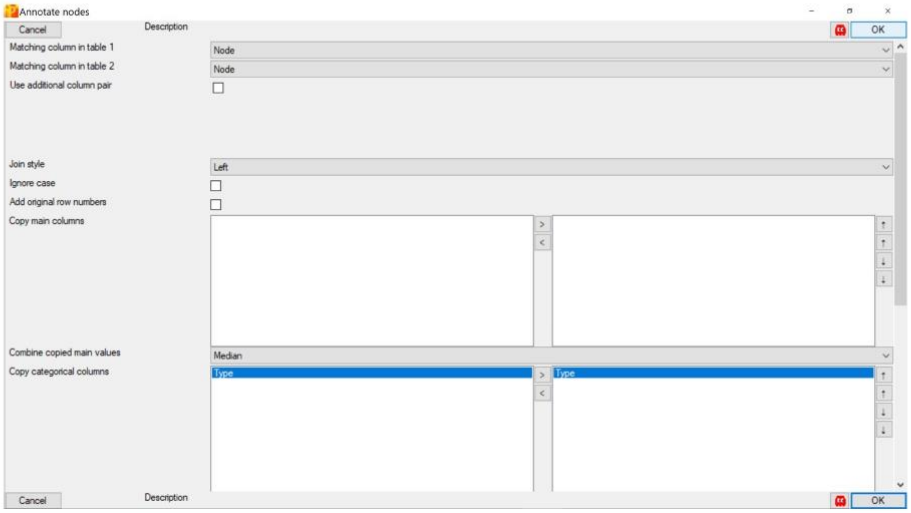
17. Select the columns corresponding to the source and target, select "Network is directed" and press "OK":



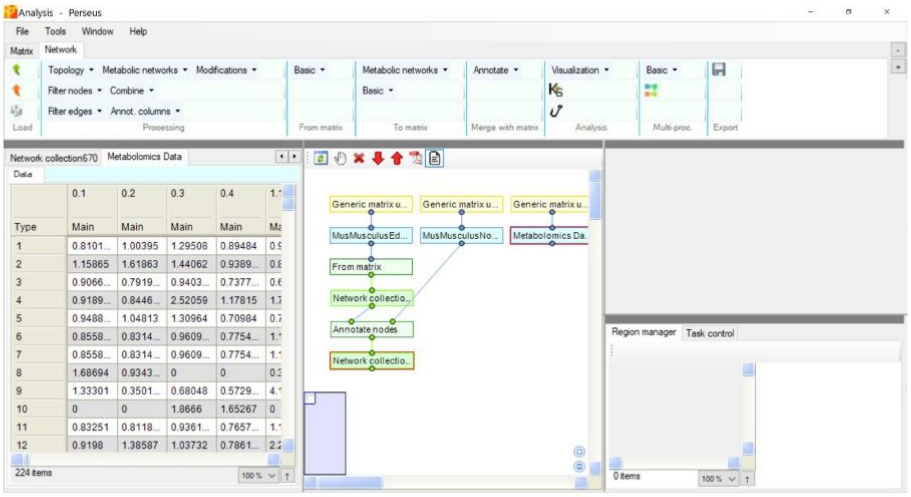
18. Select the network and the matrix (hold ctrl key to select the network and the matrix together) and on the section called “Merge with matrix”, click on “Annotate” and “Annotate nodes”:



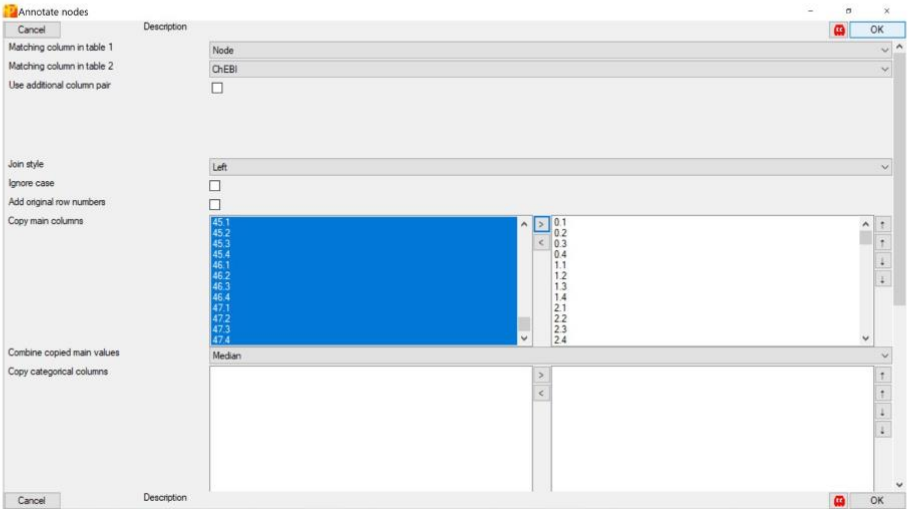
19. Select the annotations to be transferred to the network:



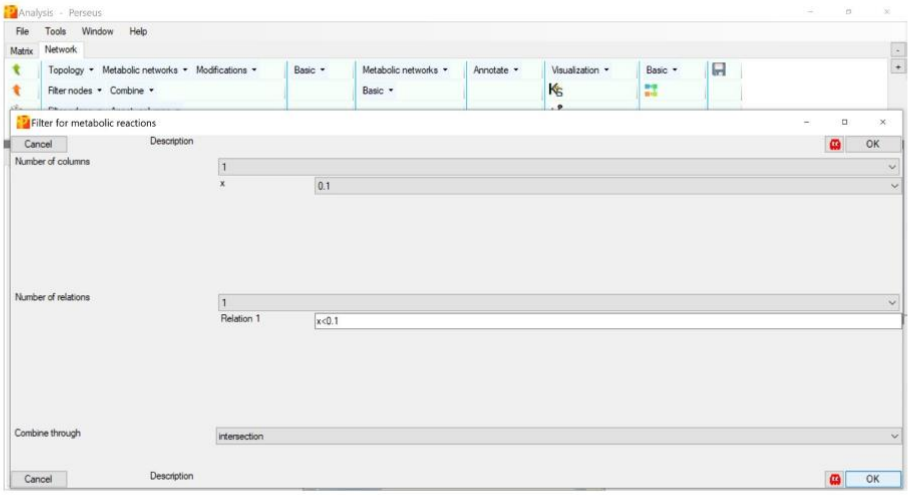
20. Step 19 can be repeated for any omics data either by using UniProt IDs or ChEBI identifiers.



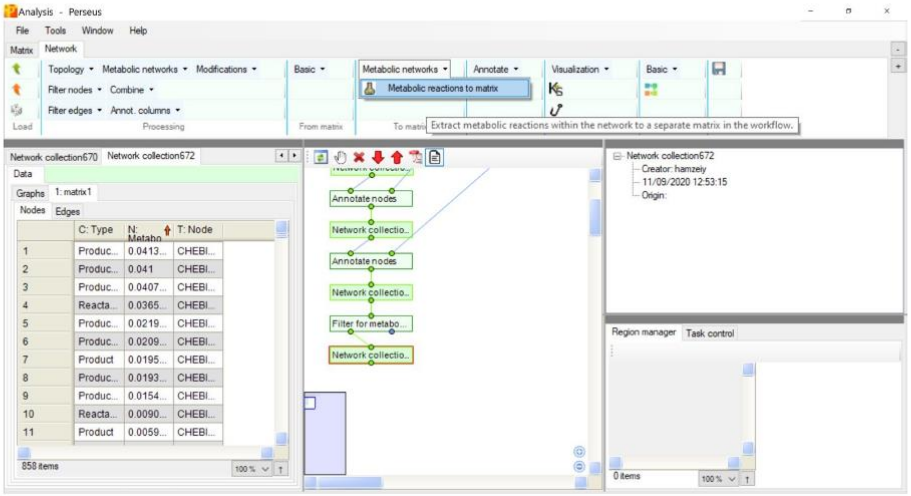
21. For example depending on the identifiers in your network and the identifiers within the omics data in question, data can be added to the network. For example metabolomics data can be added using the previously explained “Annotate nodes” function:



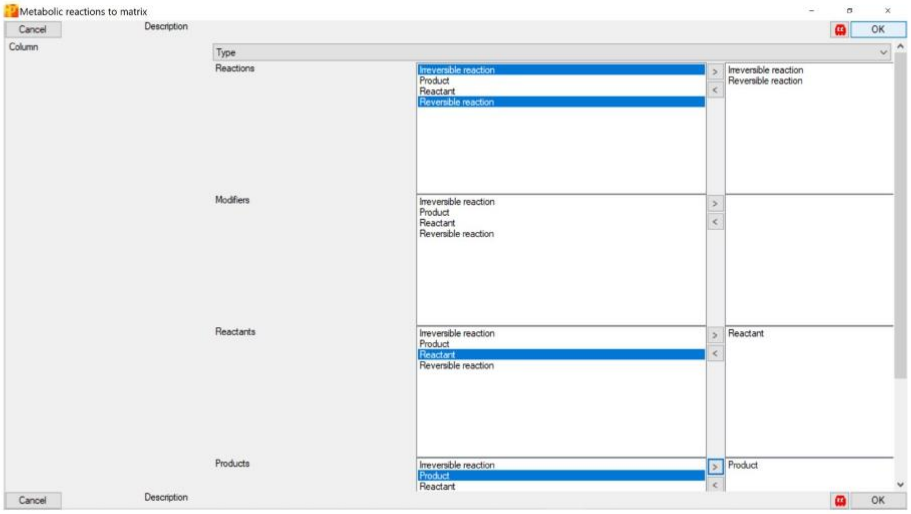
22. After annotating the network with quantitative data, one can then filter the network to retain metabolic reactions with regard to a threshold based on the quantitative data:



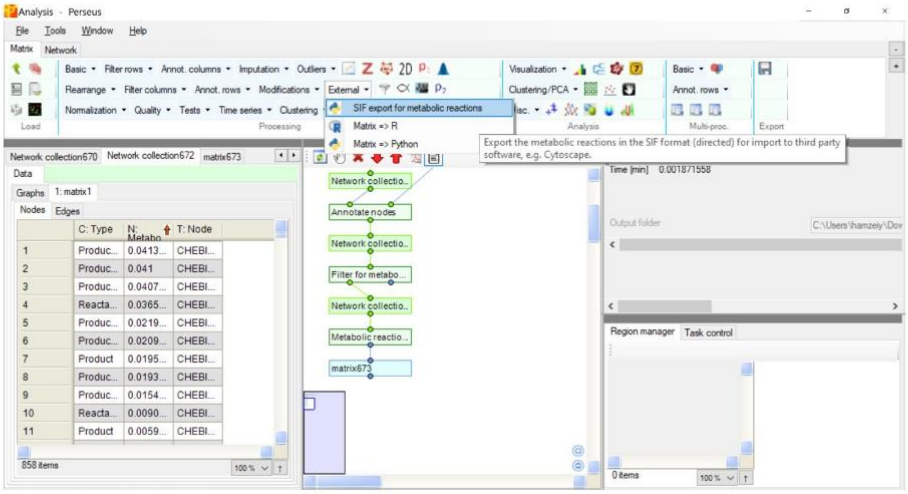
23. Once you have annotated, manipulated and filtered your network, you can convert your network to a Perseus matrix format where each reaction is depicted as a row for further analysis using other tools within Perseus:



24. In the "Metabolic reactions to matrix" dialogue box, you can choose which columns correspond to reactions, modifiers, reactants and products:



25. You can also export your network in the SIF format for further analysis in other third party software which support this generic format:


Analysis - Perseus
File Tools Window Help
Matrix Network
SIF export for metabolic reactions
Matrix => R
Matrix => Python
Export the metabolic reactions in the SIF format (directed) for import to third party software, e.g. Cytoscape.
Network collection670
Network collection672
matrix673
Data
Graphs
1: matrix1
Nodes
Edges
C: Type
T: Node
858 items
Annotate nodes
Filter for metabo...
Metabolic reactio...
Time [min] 0.001871558
Output folder
Region manager
Task control
0 items

5. Conclusion and Outlook

The rapidly evolving fields of omics, computational biology and multi-omics are fueling the data-driven revolution of the exploration of biological systems, which has had an immense impact on our understanding of the underlying mechanisms of living organisms. MaxQuant and Perseus have been continuously developed and widely accepted as trusted software platforms for the processing and analysis of shotgun proteomics datasets. These platforms house a wealth of tools and algorithms that can handle various aspects of the data and provide a user-friendly graphical interface. Expanding these platforms to various operating systems others than their native Microsoft Windows environment and their abilities in analyzing other types of proteomics data such as DIA and metabolomics datasets, provide an substantial added value as researchers are enabled to do more and with higher flexibility with software that they already use and know well.

MaxQuant is equipped with ELI for mass recalibration in metabolomics, which is the first step in developing MaxQuant as an all-in-one solution for metabolomics studies, similar to the position of MaxQuant in proteomics. Achieving higher mass accuracies in proteomics data has proven to be of paramount importance in increasing the coverage of the proteome and robustness of the approach, and similar benefits can be expected for metabolomics datasets. In addition, many of the universal functionalities of MaxQuant for LC-MS/MS data can also be readily transferred to the processing of metabolomics datasets, accelerating the development process.

Ever since the introduction of MaxQuant in 2008, it had remained a Microsoft Windows only software. With increased quantities of proteomics data becoming available, bioinformatics facilities began to include MaxQuant in their arsenal of frequently used tools. This made the need to release a Linux version of MaxQuant evident. Currently, MaxQuant runs on Windows and Linux operating systems.

Historically, MaxQuant has always been the preferred software for the analysis of DDA shotgun proteomics data, primarily due to its superior performance and ease of use. With the advancements and popularization of DIA methods for proteomics studies, the community lacked a reliable and free software for DIA data and thus, MaxQuant is now equipped with MaxDIA as a one stop DIA data processing solution. MaxDIA is capable of analyzing a wide variety of DIA data ranging from

BoxCar-DIA to ion mobility DIA data. It achieves comprehensive proteome coverage and precise quantification across many runs. Future developments of MaxDIA will focus on expanding the workflow to support PTMs; especially with respect to their correct localization within the peptide sequence.

Although Perseus was initially designed primarily for the downstream analysis of MaxQuant outputs, its popularity, flexibility and user-friendly approach for analyzing data, has made it popular for use in analyzing other omics datasets, e.g. transcriptomics. In this direction, Perseus is continuously developed to allow for various other types of data such as NGS data and biological networks. Users can now also integrate their own scripts written in Python or R and develop their own plugins in C#. With the addition of Metis, a plugin for multi-omics data analysis using metabolic networks, Perseus is now allows the user to analyze several different omics datasets together.

References

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

Aebersold, R. and Mann, M. (2016) 'Mass-spectrometric exploration of proteome structure and function', *Nature*. Nature Publishing Group, 537(7620), pp. 347–355. doi: 10.1038/nature19949.

Agrawal, A. and Choudhary, A. (2016) 'Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science', *APL Materials*. American Institute of Physics Inc., 4(5), p. 053208. doi: 10.1063/1.4946894.

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

Behjati, S. and Tarpey, P. S. (2013) 'What is next generation sequencing?', *Archives of Disease in Childhood: Education and Practice Edition*. BMJ Publishing Group, 98(6), pp. 236–238. doi: 10.1136/archdischild-2013-304340.

Bekker-Jensen, D. B. *et al.* (2017) 'An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes', *Cell Systems*. Cell Press, 4(6), pp. 587-599.e4. doi: 10.1016/j.cels.2017.05.009.

Bilbao, A. *et al.* (2015) 'Processing strategies and software solutions for data-independent acquisition in mass spectrometry', *Proteomics*. Wiley-VCH Verlag, pp. 964–980. doi: 10.1002/pmic.201400323.

Blum, B. C., Mousavi, F. and Emili, A. (2018) 'Single-platform "multi-omic" profiling: Unified mass spectrometry and computational workflows for integrative proteomics-metabolomics analysis', *Molecular Omics*, 14(5), pp. 307–319. doi: 10.1039/c8mo00136g.

Brown, K. A. *et al.* (2020) 'Top-down Proteomics: Challenges, Innovations, and Applications in Basic and Clinical Research', *Expert Review of Proteomics*. Informa UK Limited, pp. 1–15. doi: 10.1080/14789450.2020.1855982.

Büchel, F. *et al.* (2013) 'Path2Models: Large-scale generation of computational models from biochemical pathway maps', *BMC Systems Biology*, 7(1),

p. 116. doi: 10.1186/1752-0509-7-116.

Bumgarner, R. (2013) 'Overview of dna microarrays: Types, applications, and their future', *Current Protocols in Molecular Biology*, (SUPPL.101). doi: 10.1002/0471142727.mb2201s101.

Chang, J. T. (2016) 'Transcriptomics and Gene Regulation', *Transcriptomics and Gene Regulation*, 9, pp. 99–113. Available at: http://link.springer.com/10.1007/978-94-017-7450-5_4%0Ahttp://link.springer.com/10.1007/978-94-017-7450-5.

Chapman, J. D., Goodlett, D. R. and Masselon, C. D. (2014) 'Multiplexed and data-independent tandem mass spectrometry for global proteome profiling', *Mass Spectrometry Reviews*. John Wiley and Sons Inc., 33(6), pp. 452–470. doi: 10.1002/mas.21400.

Chavan, S. S., Shaughnessy, J. D. and Edmondson, R. D. (2011) 'Overview of biological database mapping services for interoperation between different "omics" datasets', *Human Genomics*, 5(6), pp. 703–708. doi: 10.1186/1479-7364-5-6-703.

Chen, C. *et al.* (2020) 'Bioinformatics methods for mass spectrometry-based proteomics data analysis', *International Journal of Molecular Sciences*, 21(8). doi: 10.3390/ijms21082873.

Chen, G., Ning, B. and Shi, T. (2019) 'Single-cell RNA-seq technologies and related computational data analysis', *Frontiers in Genetics*, 10(APR). doi: 10.3389/fgene.2019.00317.

Chen, Y. X. *et al.* (2020) 'An integrative multi-omics network-based approach identifies key regulators for breast cancer', *Computational and Structural Biotechnology Journal*, 18, pp. 2826–2835. doi: 10.1016/j.csbj.2020.10.001.

Chong, J. and Xia, J. (2017) 'Computational approaches for integrative analysis of the metabolome and microbiome', *Metabolites*, 7(4). doi: 10.3390/metabo7040062.

Clough, E. and Barrett, T. (2016) 'The Gene Expression Omnibus database', in *Methods in Molecular Biology*, pp. 93–110. doi: 10.1007/978-1-4939-3578-9_5.

Collins, B. C. *et al.* (2017) 'Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry', *Nature*

Communications. Nat ure Publishing Group, 8(1). doi: 10.1038/s41467-017-00249-5.

Cox, J. *et al.* (2011) 'Andromeda: A peptide search engine integrated into the MaxQuant environment', *Journal of Proteome Research*, 10(4), pp. 1794–1805. doi: 10.1021/pr101065j.

Cox, J. *et al.* (2014) 'Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ', *Molecular and Cellular Proteomics*, 13(9), pp. 2513–2526. doi: 10.1074/mcp.M113.031591.

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

Cox, J. and Mann, M. (2009) 'Computational Principles of Determining and Improving Mass Precision and Accuracy for Proteome Measurements in an Orbitrap', *Journal of the American Society for Mass Spectrometry*, 20(8), pp. 1477–1485. doi: 10.1016/j.jasms.2009.05.007.

Cox, J. and Mann, M. (2011) 'Quantitative, high-resolution proteomics for data-driven systems biology', *Annual Review of Biochemistry*. Annual Reviews, 80(1), pp. 273–299. doi: 10.1146/annurev-biochem-061308-093216.

Cox, J., Michalski, A. and Mann, M. (2011) 'Software lock mass by two-dimensional minimization of peptide mass errors', *Journal of the American Society for Mass Spectrometry*, 22(8), pp. 1373–1380. doi: 10.1007/s13361-011-0142-8.

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

Dahimiwal, S. M. *et al.* (2013) 'A review on high performance liquid chromatography', *International Journal of Pharmaceutical Research*. Research and Reviews, 5(3), pp. 1–6. doi: 10.22214/ijraset.2018.2098.

Dettmer, K., Aronov, P. A. and Hammock, B. D. (2007) 'Mass spectrometry-based metabolomics', *Mass Spectrometry Reviews*. NIH Public Access, pp. 51–78. doi: 10.1002/mas.20108.

Ding, L., Rath, E. and Bai, Y. (2017) 'Comparison of Alternative Splicing Junction Detection Tools Using RNASeq Data', *Current Genomics*, 18(3), pp. 268–277. doi: 10.2174/1389202918666170215125048.

Doerr, A. (2014) 'DIA mass spectrometry', *Nature Methods*, 12(1), p. 35. doi: 10.1038/nmeth.3234.

Dupree, E. J. *et al.* (2020) 'A critical review of bottom-up proteomics: The good, the bad, and the future of this field', *Proteomes*, 8(3), pp. 1–26. doi: 10.3390/proteomes8030014.

Egertson, J. D. *et al.* (2013) 'Multiplexed MS/MS for improved data-independent acquisition', *Nature Methods*. Nat Methods, 10(8), pp. 744–746. doi: 10.1038/nmeth.2528.

El-Aneed, A., Cohen, A. and Banoub, J. (2009) 'Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers', *Applied Spectroscopy Reviews*, 44(3), pp. 210–230. doi: 10.1080/05704920902717872.

Elias, J. E. and Gygi, S. P. (2007) 'Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry', *Nature Methods*, 4(3), pp. 207–214. doi: 10.1038/nmeth1019.

Fornelli, L. *et al.* (2017) 'Advancing Top-down Analysis of the Human Proteome Using a Benchtop Quadrupole-Orbitrap Mass Spectrometer', *Journal of Proteome Research*, 16(2), pp. 609–618. doi: 10.1021/acs.jproteome.6b00698.

Fuller, J. C. *et al.* (2013) 'Biggest challenges in bioinformatics', *EMBO Reports*. John Wiley & Sons, Ltd, 14(4), pp. 302–304. doi: 10.1038/embor.2013.34.

G. Marshall, A. *et al.* (2013) 'Mass Resolution and Mass Accuracy: How Much Is Enough?', *Mass Spectrometry*, 2(Special_Issue), pp. S0009–S0009. doi: 10.5702/massspectrometry.s0009.

Gauthier, J. *et al.* (2019) 'A brief history of bioinformatics', *Briefings in Bioinformatics*, 20(6), pp. 1981–1996. doi: 10.1093/bib/bby063.

Geer, L. Y. *et al.* (2004) 'Open mass spectrometry search algorithm', *Journal of Proteome Research*, 3(5), pp. 958–964. doi: 10.1021/pr0499491.

Geiger, T., Cox, J. and Mann, M. (2010a) 'Proteomic changes resulting from gene copy number variations in cancer cells', *PLoS Genetics*, 6(9), p. e1001090. doi:

10.1371/journal.pgen.1001090.

Geiger, T., Cox, J. and Mann, M. (2010b) 'Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation', *Molecular and Cellular Proteomics*. Mol Cell Proteomics, 9(10), pp. 2252–2261. doi: 10.1074/mcp.M110.001537.

Del Giacco, L. and Cattaneo, C. (2012) *Introduction to genomics*, *Methods in Molecular Biology*. doi: 10.1007/978-1-60327-216-2_6.

Giannopoulou, E. *et al.* (2019) 'Integrating next-generation sequencing in the clinical pharmacogenomics workflow', *Frontiers in Pharmacology*, 10(APR). doi: 10.3389/fphar.2019.00384.

Gillet, L. C. *et al.* (2012) 'Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis', *Molecular and Cellular Proteomics*. American Society for Biochemistry and Molecular Biology Inc., 11(6). doi: 10.1074/mcp.O111.016717.

Hale, O. J. *et al.* (2020) 'High-Field Asymmetric Waveform Ion Mobility Spectrometry and Native Mass Spectrometry: Analysis of Intact Protein Assemblies and Protein Complexes', *Analytical Chemistry*, 92(10), pp. 6811–6816. doi: 10.1021/acs.analchem.0c00649.

Hamzeiy, H. and Cox, J. (2017) 'What computational non-targeted mass spectrometry-based metabolomics can gain from shotgun proteomics', *Current Opinion in Biotechnology*, 43, pp. 141–146. doi: 10.1016/j.copbio.2016.11.014.

Haug, K. *et al.* (2013) 'MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data', *Nucleic Acids Research*. Oxford University Press, 41(D1), pp. D781–D786. doi: 10.1093/nar/gks1004.

Haug, K. *et al.* (2020) 'MetaboLights: A resource evolving in response to the needs of its scientific community', *Nucleic Acids Research*, 48(D1), pp. D440–D444. doi: 10.1093/nar/gkz1019.

Hebert, A. S. *et al.* (2018) 'Improved Precursor Characterization for Data-Dependent Mass Spectrometry', *Analytical Chemistry*. American Chemical Society, 90(3), pp. 2333–2340. doi: 10.1021/acs.analchem.7b04808.

Hein, M. Y. *et al.* (2013) 'Proteomic Analysis of Cellular Systems', in *Handbook of Systems Biology*. Elsevier, pp. 3–25. doi: 10.1016/B978-0-12-385944-0.00001-0.

Hu, Q. *et al.* (2005) 'The Orbitrap: A new mass spectrometer', *Journal of Mass Spectrometry*, pp. 430–443. doi: 10.1002/jms.856.

Huang, T. *et al.* (2012) 'Protein inference: A review', *Briefings in Bioinformatics*, 13(5), pp. 586–614. doi: 10.1093/bib/bbs004.

Hurd, P. J. and Nelson, C. J. (2009) 'Advantages of next-generation sequencing versus the microarray in epigenetic research', *Briefings in Functional Genomics and Proteomics*, 8(3), pp. 174–183. doi: 10.1093/bfgp/elp013.

Ingolia, N. T. (2014) 'Ribosome profiling: New views of translation, from single codons to genome scale', *Nature Reviews Genetics*, 15(3), pp. 205–213. doi: 10.1038/nrg3645.

Kandpal, R. P., Saviola, B. and Felton, J. (2009) 'The era of 'omics unlimited', *BioTechniques*. Future Science Ltd London, UK , pp. 351–355. doi: 10.2144/000113137.

Kanter, I. and Kalisky, T. (2015) 'Single cell transcriptomics: Methods and applications', *Frontiers in Oncology*, 5(FEB). doi: 10.3389/fonc.2015.00053.

Karahalil, B. (2016) 'Overview of Systems Biology and Omics Technologies', *Current Medicinal Chemistry*, 23(37), pp. 4221–4230. doi: 10.2174/0929867323666160926150617.

Katajamaa, M. and Orešič, M. (2005) 'Processing methods for differential analysis of LC/MS profile data', *BMC Bioinformatics*, 6. doi: 10.1186/1471-2105-6-179.

Kim, M. and Tagkopoulos, I. (2018) 'Data integration and predictive modeling methods for multi-omics datasets', *Molecular Omics*, 14(1), pp. 8–25. doi: 10.1039/c7mo00051k.

Kind, T. and Fiehn, O. (2006) 'Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm', *BMC Bioinformatics*, 7. doi: 10.1186/1471-2105-7-234.

Koboldt, D. C. *et al.* (2013) 'XThe next-generation sequencing revolution and

its impact on genomics', *Cell*, 155(1), p. 27. doi: 10.1016/j.cell.2013.09.006.

Kodama, Y., Shumway, M. and Leinonen, R. (2012) 'The sequence read archive: Explosive growth of sequencing data', *Nucleic Acids Research*, 40(D1), pp. D54–D56. doi: 10.1093/nar/gkr854.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Lange, E. *et al.* (2008) 'Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements', *BMC Bioinformatics*, 9. doi: 10.1186/1471-2105-9-375.

Larsen, M. R. *et al.* (2006) 'Analysis of posttranslational modifications of proteins by tandem mass spectrometry', *BioTechniques*, 40(6), pp. 790–798. doi: 10.2144/000112201.

Levy, S. E. and Myers, R. M. (2016) 'Advancements in Next-Generation Sequencing', *Annual Review of Genomics and Human Genetics*, 17, pp. 95–115. doi: 10.1146/annurev-genom-083115-022413.

Libiseller, G. *et al.* (2015) 'IPO: A tool for automated optimization of XCMS parameters', *BMC Bioinformatics*, 16(1). doi: 10.1186/s12859-015-0562-8.

Luscombe, N. M., Greenbaum, D. and Gerstein, M. (2001) 'What is bioinformatics? A proposed definition and overview of the field', *Methods of Information in Medicine*. Schattauer GmbH, 40(4), pp. 346–358. doi: 10.1055/s-0038-1634431.

MacKlin, P. (2019) 'Key challenges facing data-driven multicellular systems biology', *GigaScience*, 8(10). doi: 10.1093/gigascience/giz127.

Mailman, M. D. *et al.* (2007) 'The NCBI dbGaP database of genotypes and phenotypes', *Nature Genetics*, 39(10), pp. 1181–1186. doi: 10.1038/ng1007-1181.

Makarov, A. (2000) 'Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis', *Analytical Chemistry*. Anal Chem, 72(6), pp. 1156–1162. doi: 10.1021/ac991131p.

Makarov, A. *et al.* (2006) 'Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer', *Journal of the American Society for Mass Spectrometry*, 17(7), pp. 977–982. doi: 10.1016/j.jasms.2006.03.006.

Marshall, A. G. and Hendrickson, C. L. (2008) 'High-resolution mass spectrometers', *Annual Review of Analytical Chemistry*. Annu Rev Anal Chem (Palo Alto Calif), pp. 579–599. doi: 10.1146/annurev.anchem.1.031207.112945.

Marshall, C. R. (2006) 'Mass Spectrometry: Bottom-Up or Top-Down?', *Science*, 314(5796), pp. 66–67.

Marx, V. (2013) 'Targeted proteomics', *Nature Methods*. Nat Methods, 10(1), pp. 19–22. doi: 10.1038/nmeth.2285.

Masselon, C. *et al.* (2000) 'Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures', *Analytical Chemistry*. American Chemical Society , 72(8), pp. 1918–1924. doi: 10.1021/ac991133+.

de Matos, P. *et al.* (2010) 'Chemical Entities of Biological Interest: an update', *Nucleic acids research*, 38(Database issue), pp. D249–D254. doi: 10.1093/nar/gkp886.

Michalski, A. *et al.* (2011) 'Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer', *Molecular and Cellular Proteomics*, 10(9), p. M111.011015. doi: 10.1074/mcp.M111.011015.

Michalski, A., Cox, J. and Mann, M. (2011) 'More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS', *Journal of Proteome Research*. American Chemical Society, 10(4), pp. 1785–1793. doi: 10.1021/pr101060v.

Miladinović, S. M. *et al.* (2012) 'On the utility of isotopic fine structure mass spectrometry in protein identification', *Analytical Chemistry*, 84(9), pp. 4042–4051. doi: 10.1021/ac2034584.

Mishra, N. (2010) *Introduction to Proteomics: Principles and Applications*, *Introduction to Proteomics: Principles and Applications*. doi: 10.1002/9780470603871.

Nagaraj, N. *et al.* (2011) 'Deep proteome and transcriptome mapping of a human cancer cell line', *Molecular Systems Biology*. Mol Syst Biol, 7, p. 548. doi: 10.1038/msb.2011.81.

Navarro, P. *et al.* (2016) 'A multicenter study benchmarks software tools for label-free proteome quantification', *Nature Biotechnology*. Nature Publishing Group, 34(11), pp. 1130–1136. doi: 10.1038/nbt.3685.

Olsen, J. V. *et al.* (2005) 'Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap', *Molecular and Cellular Proteomics*, 4(12), pp. 2010–2021. doi: 10.1074/mcp.T500030-MCP200.

Orchard, S. *et al.* (2014) 'The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1115.

Oughtred, R. *et al.* (2020) 'The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions', *Protein Science*. doi: 10.1002/pro.3978.

Panchaud, A. *et al.* (2009) 'Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean', *Analytical Chemistry*. American Chemical Society, 81(15), pp. 6481–6488. doi: 10.1021/ac900888s.

Panchaud, A. *et al.* (2011) 'Faster, quantitative, and accurate precursor acquisition independent from ion count', *Analytical Chemistry*. Anal Chem, 83(6), pp. 2250–2257. doi: 10.1021/ac103079q.

Paša-Tolić, L. *et al.* (2004) 'Proteomic analyses using an accurate mass and time tag strategy', *BioTechniques*, 37(4), pp. 621–639. doi: 10.2144/04374rv01.

Piazza, I. *et al.* (2018) 'A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication', *Cell*. Elsevier, 172(1–2), pp. 358-372.e23. doi: 10.1016/j.cell.2017.12.006.

Pluskal, T. *et al.* (2010) 'MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data', *BMC Bioinformatics*, 11. doi: 10.1186/1471-2105-11-395.

Rauluseviciute, I., Drabløs, F. and Rye, M. B. (2019) 'DNA methylation data by sequencing: Experimental approaches and recommendations for tools and pipelines for data analysis', *Clinical Epigenetics*, 11(1). doi: 10.1186/s13148-019-0795-x.

Röst, H. L. *et al.* (2014) 'OpenSWATH enables automated, targeted analysis

of data-independent acquisition MS data', *Nature Biotechnology*. Nature Publishing Group, pp. 219–223. doi: 10.1038/nbt.2841.

Rotroff, D. M. and Motsinger-Reif, A. A. (2016) 'Embracing Integrative Multiomics Approaches', *International Journal of Genomics*, 2016. doi: 10.1155/2016/1715985.

Scigelova, M. *et al.* (2011) 'Fourier transform mass spectrometry', *Molecular and Cellular Proteomics*. American Society for Biochemistry and Molecular Biology. doi: 10.1074/mcp.M111.009431.

Shahaf, N. *et al.* (2013) 'Constructing a mass measurement error surface to improve automatic annotations in liquid chromatography/mass spectrometry based metabolomics', *Rapid Communications in Mass Spectrometry*, 27(21), pp. 2425–2431. doi: 10.1002/rcm.6705.

Sinitcyn, P. *et al.* (2018) 'MaxQuant goes Linux', *Nature Methods*, 15(6), p. 401. doi: 10.1038/s41592-018-0018-y.

Sinitcyn, P., Rudolph, J. D. and Cox, J. (2018) 'Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data', *Annual Review of Biomedical Data Science*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, California 94303-0139, USA, 1(1), pp. 207–234. doi: 10.1146/annurev-biodatasci-080917-013516.

Smith, R. *et al.* (2014) 'Proteomics, lipidomics, metabolomics: A mass spectrometry tutorial from a computer scientist's point of view', *BMC Bioinformatics*, 15. doi: 10.1186/1471-2105-15-S7-S9.

Stringer, K. A. *et al.* (2016) 'Metabolomics and its application to acute lung diseases', *Frontiers in Immunology*, 7(FEB), p. 44. doi: 10.3389/fimmu.2016.00044.

Subramanian, I. *et al.* (2020) 'Multi-omics Data Integration, Interpretation, and Its Application', *Bioinformatics and Biology Insights*, 14. doi: 10.1177/1177932219899051.

Szklarczyk, D. *et al.* (2019) 'STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets', *Nucleic Acids Research*, 47(D1), pp. D607–D613. doi: 10.1093/nar/gky1131.

Tautenhahn, R., Bottcher, C. and Neumann, S. (2008) 'Highly sensitive feature detection for high resolution LC/MS', *BMC Bioinformatics*, 9. doi: 10.1186/1471-2105-9-504.

Toby, T. K., Fornelli, L. and Kelleher, N. L. (2016) 'Progress in Top-Down Proteomics and the Analysis of Proteoforms', *Annual Review of Analytical Chemistry*, 9, pp. 499–519. doi: 10.1146/annurev-anchem-071015-041550.

Trapnell, C. (2015) 'Defining cell types and states with single-cell genomics', *Genome Research*, 25(10), pp. 1491–1498. doi: 10.1101/gr.190595.115.

Tyanova, S. *et al.* (2016) 'The Perseus computational platform for comprehensive analysis of (prote)omics data', *Nature Methods*, 13(9), pp. 731–740. doi: 10.1038/nmeth.3901.

Tyanova, S., Temu, T. and Cox, J. (2016) 'The MaxQuant computational platform for mass spectrometry-based shotgun proteomics', *Nature Protocols*, 11(12), pp. 2301–2319. doi: 10.1038/nprot.2016.136.

Uppal, K. *et al.* (2013) 'XMSanalyzer: Automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data', *BMC Bioinformatics*, 14(1). doi: 10.1186/1471-2105-14-15.

Uppal, K. *et al.* (2016) 'Computational Metabolomics: A Framework for the Million Metabolome', *Chemical Research in Toxicology*, 29(12), pp. 1956–1975. doi: 10.1021/acs.chemrestox.6b00179.

Venable, J. D. *et al.* (2004) 'Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra', *Nature Methods*. Nat Methods, 1(1), pp. 39–45. doi: 10.1038/nmeth705.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: A revolutionary tool for transcriptomics', *Nature Reviews Genetics*, 10(1), pp. 57–63. doi: 10.1038/nrg2484.

Weckwerth, W. (2007) *Metabolomics Methods and Protocols, Synthesis*. doi: 10.1007/978-1-59745-244-1.

Wiśniewski, J. R. *et al.* (2014) 'A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards', *Molecular and Cellular Proteomics*, 13(12), pp. 3497–3506. doi: 10.1074/mcp.M113.037309.

Wolf-Yadlin, A., Hu, A. and Noble, W. S. (2016) 'Technical advances in proteomics: New developments in data-independent acquisition', *F1000Research*. Faculty of 1000 Ltd, 5. doi: 10.12688/f1000research.7042.1.

Wolters, D. A., Washburn, M. P. and Yates, J. R. (2001) 'An automated multidimensional protein identification technology for shotgun proteomics', *Analytical Chemistry*, 73(23), pp. 5683–5690. doi: 10.1021/ac010617e.

Wu, L. and Han, D. K. (2006) 'Overcoming the dynamic range problem in mass spectrometry-based shotgun proteomics', *Expert Review of Proteomics*, 3(6), pp. 611–619. doi: 10.1586/14789450.3.6.611.

Yu, T. *et al.* (2009) 'apLCMS-adaptive processing of high-resolution LC/MS data', *Bioinformatics*, 25(15), pp. 1930–1936. doi: 10.1093/bioinformatics/btp291.

Zhang, F. *et al.* (2020) 'Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020', *Proteomics*, 20(17–18). doi: 10.1002/pmic.201900276.

Zhang, J. *et al.* (2009) 'Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry', *Current Genomics*, 10(6), pp. 388–401. doi: 10.2174/138920209789177638.

Zhang, Y. *et al.* (2013) 'Protein analysis by shotgun/bottom-up proteomics', *Chemical Reviews*. Chem Rev, 113(4), pp. 2343–2394. doi: 10.1021/cr3003533.

List of Figures

List of Symbols, Acronyms and Abbreviations

NGS	Next Generation Sequencing
LC-MS/MS	Liquid Chromatography coupled to Tandem Mass Spectrometry
DIA	Data Independent Acquisition
Da	Dalton
PTM	Post Translational Modification
m/z	Mass to Charge Ratio
FWHM	Full Width Half Maximum
FTMS	Fourier Transform Ion Cyclotron Resonance Mass Spectrometer
PSM	Peptide Spectrum Match
FDR	False Discover Rate
LFQ	Label Free Quantification
DDA	Data Dependent Acquisition
ELI	Easy Library Implementation

Acknowledgements

I would like to extend my deepest gratitude to Dr. Jürgen Cox, for his guidance and feedback throughout this work. It has been a pleasure to have the opportunity to work with such a great mind.

I would also like to thank Prof. Dr. Christoph Turck, for taking the time to be there for me as a member of my thesis advisory committee, for his on point comments and suggestion, fruitful discussions, and for being my internal supervisor at LMU.

I am also very grateful to have had the chance to work with Prof. Dr. Maria Robles for a part of this work, from which I have learnt a great deal.

Many thanks to Dr. Gabi Kastenmüller, for her help in my thesis advisory committee.

Many special thanks go to Dr. Christoph Wichmann, who has not only been a very valuable colleague from which I have learnt many things, and of great help by proofreading this work, but also a very dear friend, whose support, guidance and advice has significantly enhanced my time in the lab.

I am happy and thankful to Dr. Sule Yilmaz-Rumpf for help with proofreading of this work and as a friend in the lab to enjoy Turkish coffee with, and have fruitful discussions about the various aspects of our projects.

I also want to thank the entire lab, for creating a pleasant work atmosphere. Special thanks to Daniela who has always been helpful in various aspects of programming (not to forget the delicious coffee), Assa for his help to solve any problems within the lab (also for his pleasant company), Favio for being a great friend and Peli for her special sense of humor, which I enjoy very much.

Many thanks to the IMPRS office, for their help with the many different stages of the PhD from the very first day until now. I really appreciate it.

Thank you Ece, for all the good times and memories.

I would also like to thank all my friends, who have always been there for me and have supported me throughout this work.

As always last, but certainly not least, I would like to thank my amazing parents, who have been paramount in their support and guidance. Without them, I cannot imagine having been able to achieve any of my goals and aspirations. I am forever thankful to them.

Curriculum Vitae

Contact

hamidhamzeiy@gmail.com

+49 176 471 34668

Top Skills

- Bioinformatics
- Programming (Python & C#)
- Artificial Intelligence
- Proteomics
- Metabolomics
- Genomics

Languages

- Azerbaijani (Native or Bilingual)
- English (Native or Bilingual)
- German (Intermediate - B1)
- Persian (Native or Bilingual)
- Turkish (Native or Bilingual)

Certifications

- DTZ German -B1
- Good Manufacturing Practices (GMP)

Honors and Awards

- İnan Kıraç Fellowship
- IUGEN Poster Competition - First Prize
- Erasmus Scholarship for a summer internship in Europe
- Boğaziçi University Research Fund

Volunteer Experience

- Running Group Trainer at Münchner Bündnis gegen Depression e.V. (2020 -Present)
- Web Administrator at SymBioSE (2012 - 2017)

Hamid Hamzeiy

PhD Student at Max Planck Institute of Biochemistry

Work Experience

Genomize

- Bioinformatician, Contract (full-time)

Back-end data anlysis pipeline & feature development (Python)

January 2016 - August 2016 (8 months)

Istanbul, Turkey

JIBtools

- Section Editor (miRNA data analysis tools)

December 2013 - September 2017 (3 years 10 months)

Bielefeld, Germany

Istituto Auxologico Italiano

- Visiting Researcher (wet-lab & microarray data analysis)

August 2014 - September 2014 (2 months)

Milan, Italy

IZTECH Student Council

- Vice President (general budget management & organization of student events and festivals)

November 2012 - February 2014 (1 year 4 months)

Izmir, Turkey

IEEE IZTECH Student Branch

- Engineering in Medicine and Biology Society Chair

May 2012 - October 2012 (6 months)

Izmir, Turkey

Bielefeld University

- Visiting Bioinformatician (miRNA data analysis & visualization)

June 2012 - September 2012 (4 months)

Bielefeld, Germany

Publications

- MaxQuant goes Linux
- What computational non-targeted mass spectrometry-based metabolomics can gain from shotgun proteomics
- Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis
- Computational methods for microRNA target prediction
- Search for SCA2 blood RNA biomarkers highlights Ataxin-2 as strong modifier of the mitochondrial factor PINK1 levels
- Visualization and Analysis of MicroRNAs within KEGG Pathways using VANESA
- Revisiting the complex architecture of ALS in Turkey: Expanding genotypes, shared phenotypes, molecular networks, and a public variant database
- Characterization of the c9orf72 GC-rich low complexity sequence in two cohorts of Italian and Turkish ALS cases
- Can MiRBase provide positivedata for machine learning for the detection of MiRNA hairpins?
- Elevated global DNA methylation is not exclusive to amyotrophic lateral sclerosis and is also observed in spinocerebellar ataxia types 1 and 2

Education

IMPRS-LS | International Max Planck Research School for Molecular Life Sciences

- Doctor of Philosophy (PhD), Bioinformatics · (2016 - 2021)

Development of MaxQuant and Perseus (C#)

Max Planck Institute of Biochemistry
Computational Systems Biochemistry

Boğaziçi University

- MSc, Molecular Biology and Genetics · (2014 - 2016)

Experimental biology and genomics data analysis (Python)

Neurodegeneration Research Laboratory (NDAL)

İzmir Yüksek Teknoloji Enstitüsü

- BSc, Molecular Biology and Genetics · (2009 - 2014)

Bioinformatics Group (JLab)

References

Dr. Juergen Cox
Max Planck Institute of Biochemistry
Computational Systems Biochemistry
cox@biochem.mpg.de

Dr. Ersen Kavak
Genomize Inc.
ersen@genomize.com

Prof. Dr. Nazli Basak
Koc University Research Center for Translational Medicine
Neurodegeneration Research Laboratory (NDAL)
nbasak@ku.edu.tr