Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Development and Application of Software and Algorithms for Network Approaches to Proteomics Data Analysis

Jan Daniel Rudolph

aus
Tübingen, Deutschland

2019

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Professor Dr. Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 4. März 2019

Jan Daniel Rudolph

Dissertation eingereicht am:	17.01.2019
1. Gutachter:	Prof. Dr. Matthias Mann
2. Gutachter:	Prof. Dr. Jürgen Cox
Mündliche Prüfung am:	20.02.2019

Summary

The cells making up all living organisms integrate external and internal signals to carry out the functions of life. Dysregulation of signaling can lead to a variety of grave diseases, including cancer [Slamon et al., 1987]. In order to understand signal transduction, one has to identify and characterize the main constituents of cellular signaling cascades. Proteins are involved in most cellular processes and form the major class of biomolecules responsible for signal transduction. Post-translational modifications (PTMs) of proteins can modulate their enzymatic activity and their protein-protein interactions (PPIs) which in turn can ultimately lead to changes in protein expression. Classical biochemistry has approached the study of proteins, PTMs and interaction from a reductionist view. The abundance, stability and localization of proteins was studied one protein at a time, following the one gene-one protein-one function paradigm [Beadle and Tatum, 1941]. Pathways were considered to be linear, where signals would be transmitted from a gene to proteins, eventually resulting in a specific phenotype. Establishing the crucial link between genotype and phenotype remains challenging despite great advances in omics technologies, such as liquid chromatography (LC)-mass spectrometry (MS) that allow for the system-wide interrogation of proteins.

Systems and network biology [Barabási and Oltvai, 2004, Bensimon et al., 2012, Jørgensen and Locard-Paulet, 2012, Choudhary and Mann, 2010] aims to transform modern biology by utilizing omics technologies to understand and uncover the various complex networks that govern the cell. The first detected large-scale biological networks have been found to be highly structured and non-random [Albert and Barabási, 2002]. Furthermore, these are assembled from functional and topological modules. The smallest topological modules are formed by the direct physical interactions within protein-protein and protein-RNA complexes. These molecular machines are able to perform a diverse array of cellular functions, such as transcription and degradation [Alberts, 1998]. Members of functional modules are not required to have a direct physical interaction. Instead, such modules also include proteins with temporal co-regulation throughout the cell cycle [Olsen et al., 2010], or following the circadian day-night rhythm [Robles et al., 2014]. The signaling pathways that make up the cellular network [Jordan et al., 2000] are assembled from a hierarchy of these smaller modules [Barabási and Oltvai, 2004]. The regulation of these modules through dynamic

rewiring enables the cell to respond to internal an external stimuli.

The main challenge in network biology is to develop techniques to probe the topology of various biological networks, to identify topological and functional modules, and to understand their assembly and dynamic rewiring. LC-MS has become a powerful experimental platform that addresses all these challenges directly [Bensimon et al., 2012], and has long been used to study a wide range of biomolecules that participate in the cellular network. The field of proteomics in particular, which is concerned with the identification and characterization of the proteins in the cell, has been revolutionized by recent technological advances in MS. Proteomics experiments are used not only to quantify peptides and proteins, but also to uncover the edges of the cellular network, by screening for physical PPIs in a global [Hein et al., 2015] or condition specific manner [Kloet et al., 2016]. Crucial for the interpretation of the large-scale data generated by MS experiments is the development of software tools that aid researchers in translating raw measurements into biological insights. The MaxQuant and Perseus platforms were designed for this exact purpose.

The aim of this thesis was to develop software tools for the analysis of MS-based proteomics data with a focus on network biology and apply the developed tools to study cellular signaling. The first step was the extension of the Perseus software with network data structures and activities. The new network module allows for the side-by-side analysis of matrices and networks inside an interactive workflow and is described in article 1. We subsequently apply the newly developed software to study the circadian phosphoproteome of cortical synapses (see article 2). In parallel we aimed to improve the analysis of large datasets by adapting the previously Windows-only MaxQuant software to the Linux operating system, which is more prevalent in high performance computing environments (see article 3).

Contents

List of Figures

Chapter 1

Introduction

1.1 Mass spectrometry-based proteomics

The shotgun (bottom-up) approach has been established as a generic and flexible workflow for MS-based proteomics. By measuring peptides instead of intact proteins, challenges in the analysis of intact proteins are circumvented [Zhang et al., 2013]. A typical workflow begins by sample preparation, optional protein or peptide fractionation and enrichment (see Figure 1.2), high performance liquid chromatography (HPLC), and MS acquisition (see Figure 1.1). Finally, a computational analysis of the acquired data is required to identify and quantify the peptides, the proteins and their PTMs in the sample.

1.1.1 Sample preparation

The first step of the workflow is the extraction of protein material from the sample by cell or tissue lysis, followed by the enzymatic digestion of the proteins into peptides (see Figure 1.2). The most popular restriction enzyme is trypsin, which specifically cleaves C-terminal after arginine and lysine. Tryptic peptides have a convenient length distribution and favorable charge for MS. Historically, proteins where digested 'in-gel' after separating them on a SDS polyacrylamide gel [Shevchenko et al., 1996]. More recently, 'in-solution' digestion became the method of choice, especially in combination with HPLC [Wiśniewski et al., 2009, Kulak et al., 2014].

In order to reduce the sample complexity a number of different offline and online fractionation techniques can be employed prior to MS analysis. One-dimensional polyacrylamide gel electrophoresis (1D-PAGE) can be used to separate peptides according

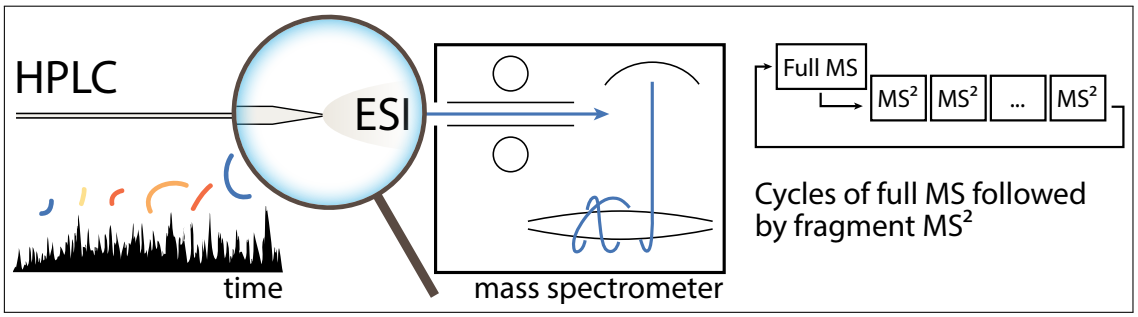


Figure 1.1: Overview over a typical LC-MS workflow. The sample elutes from the HPLC and is injected into the mass spectrometer after ESI. The machine follows a predefined acquisition strategy along its duty cycle, which combines scanning the injected ions and their fragment products by MS and MS^2 scans. Adapted from [Hein et al., 2013].

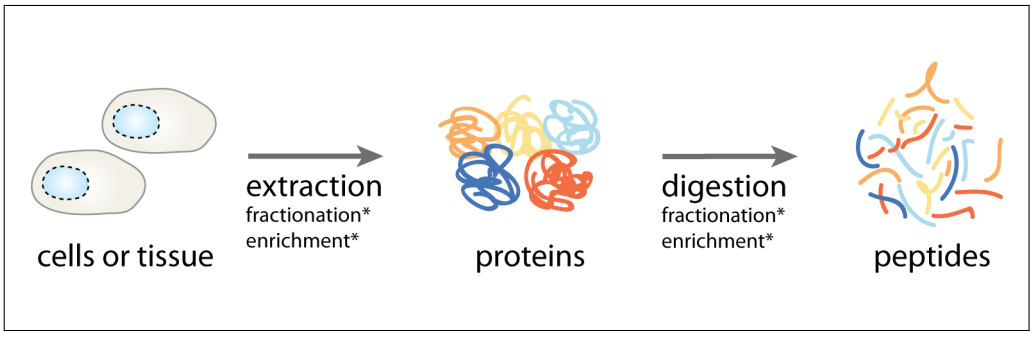


Figure 1.2: In the shotgun proteomics workflow proteins are extracted from the sample of interest and digested into peptides prior to MS analysis. Adapted from [Hein et al., 2013].

to their mass, and is easily combined with 'in-gel' digestion. Samples which are 'in-solution' are analyzed by HPLC. In HPLC systems the peptides differentially interact with the stationary phase of strong cation exchange (SCX) or reversed-phase (RP) chromatography columns due to their different physiochemical properties [Wolters et al., 2001]. RP chromatography is based on the hydrophobic interaction between the peptide and the C18-silica of the column. By applying a pH gradient to the mobile phase all peptides can be eluted from the column over the course of a MS run. Optimal chromatographic resolution can be obtained by increasing the length of the column and reducing its diameter. However, these alterations lead to increased backpressure on the HPLC system and reduced ionization efficiency [Jorgenson, 2010].

1.1.2 The mass spectrometer

The three central parts that make up the mass spectrometer and define its main characteristics are the ion source, the mass analyzer, and the detector. An ion source is required for the production of ionized particles which subsequently enter the high vacuum of the mass spectrometer. The soft-ionization technique ESI [Fenn et al., 1989] enables the analysis of intact proteins and peptides from solution, which makes it attractive for LC-MS analysis. Alternative approaches, such as matrix-assisted laser desorption/ionization (MALDI) [Karas and Hillenkamp, 1988] create ions by pulsing the sample loaded onto a solid matrix with a laser.

The mass-to-charge ratio m/z of the injected ions is measured in the mass analyzer (see Table 1.1). Beam-type analyzers include the quadrupole, in which the ions are guided through two pairs of electrodes, and the time-of-flight (TOF) analyzer, where the ions fly through a drift region that separates them according to their mass. Beam-type analyzers are characterized by their simplicity and dynamic range while having lower resolution and mass accuracy. Trapping mass analyzers include linear ion traps, Fourier transform ion cyclotrons (FTICRs) and the Orbitrap. The transients of ions oscillating inside the trapping analyzer are recorded and transformed into m/z values using Fourier analysis [Scigelova et al., 2011]. The Orbitrap has many favourable characteristics for lower-mass peptide analytes, such as very high resolution and mass accuracy [Zubarev and Makarov, 2013]. Many modern mass spectrometers, such as the Q-Exactive HF [Scheltema et al., 2014] operate multiple mass analyzers in tandem. Quadrupoles are used for the selection of ions within a specified m/z range and traps

	Quadrupole	TOF	Ion trap	FTICR	Orbitrap
Resolution	medium	high	medium	very high	very high
Accuracy	low	high	low	very high	very high
Sensitivity	very high	medium	very high	medium	very high
Dynamic range	very high	very high	medium	high	high
Speed	medium	very high	very high	medium	very high
Simplicity	very high	high	very high	low	medium

Table 1.1: Characteristics of mass analyzers commonly used in MS

are often utilized for the accumulation of ions prior to mass analysis (see Figure 1.3). Finally, the ions reach the detector which counts the number of ions observed at each m/z value.

Mass spectrometers can be operated in a number of different acquisition modes which determine the succession of full (MS^1) and fragment (MS^2) scans during a measurement run (see Figure 1.1). In targeted mode, the mass spectrometer is configured to target a predefined set of masses, aiming for the highest possible quantitative accuracy and reproducibility [Marx, 2013]. In contrast, data-dependent acquisision (DDA) relies on the observed peaks on the MS^1-level to decide which ions will be subsequently isolated, fragmented and sent for MS^2 analysis. The goal of the MS^2 analysis is to sequence the peptide by measuring the fragment ion series. To this end, fragmentation energies are optimized to induce a single peptide backbone breakage that gives rise to a set of complementary fragment ions. Time constraints do not allow for the exhaustive sequencing of peptides. Instead, a common strategy is the selection of the n most intense peaks for MS^2 [Mann et al., 2001]. With advances in instrumentation and software, data-independent acquisition (DIA) [Gillet et al., 2012] has emerged as an alternative to DDA for proteomic analysis. After acquiring the MS^1 scan, the entire mass range is segmented into overlapping windows. Subsequently, each mass window is fragmented and a fragment scan is obtained, regardless of the measured MS^1 information.

1.1.3 Quantitative Proteomics

While historically the mass spectrometer was mainly used for protein identification, the development of quantitative MS enabled a wide range of analytical techniques for MS-

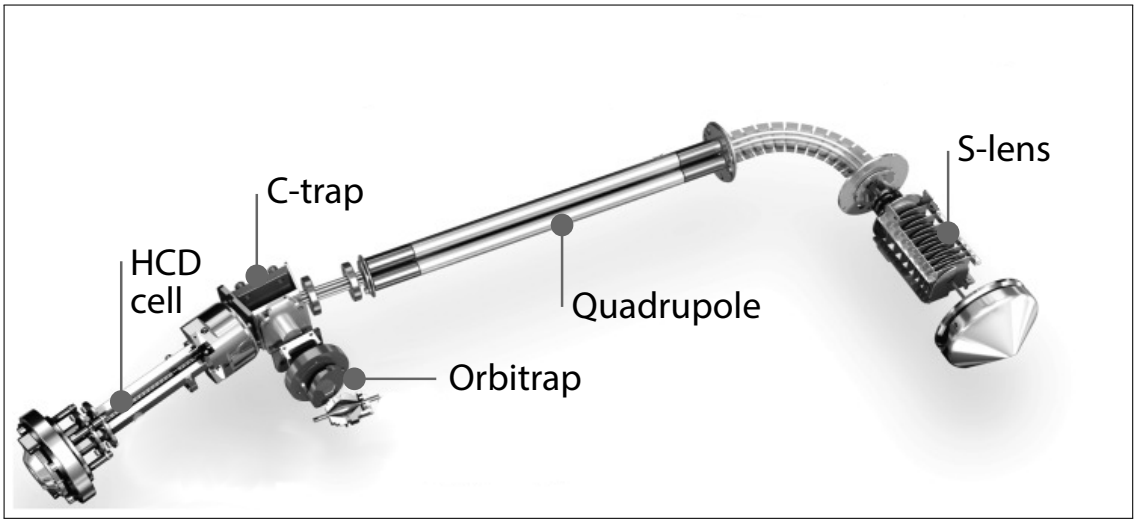


Figure 1.3: Component layout of the Q Exactive HF mass spectrometer. Ions entering the machine are focussed and filtered in the S-lens and the consecutive optics. The Q Exactive HF then combines a quadrupole for ion selection, C-trap for ion accumulation, HCD cell for fragmentation and Orbitrap for mass analysis.

based proteomics. MS is not inherently quantitative due to the effect that the different physiochemical properties of the peptides have on their behavior in the mass spectrometer. For example, differences in ionization efficiency are reflected in the measured intensities, thus making direct comparisons difficult. Label-free approaches therefore rely on computational schemes for accurate quantification of the observed intensities [Cox et al., 2014]. By labeling peptides metabolically or chemically, the introduced mass shift adds a directly measurable quantitative dimension to the experiment. Stable isotope labeling by amino acids in cell culture (SILAC) [Ong et al., 2002, Blagoev et al., 2004] exploits the incorporation of heavy or medium lysine and arginine into the proteome. The digested tryptic peptides will contain at least one labeled amino acid and are therefore distinguishable from their unlabeled counterparts. In MS analysis, SILAC triplets can be detected and their intensities compared. Alternatively, isobaric chemical labeling reagents such as tandem mass tag (TMT) [Thompson et al., 2003]. The isobaric label is constructed of up to 11 reporters with distinct mass and a corresponding balancer [Werner et al., 2014]. While isobarically labeled peptides are indistinguishable on the MS^1 level, on the MS^2 level a reporter ion fragment can be observed for each channel.

SILAC labeling provides the highest quantification accuracy but is limited in the number of channels. With up to 11 channels, TMT has increased multiplexing capabilities but an accurate MS^2 quantification without ratio compression introduced by co-eluting peptides requires specialized mass spectrometers [Savitski et al., 2013]. Compared to the label-free approach, TMT showed higher precision and fewer missing values [O'Connell et al., 2018], however, only the label-free approach can scale to an arbitrary number of samples

1.2 Computational mass spectrometry

In the following review a broad introduction to the computational aspects of MS-based proteomics is presented. It covers the identification and quantification of peptides, proteins and PTMs, as well the statistical downstream data analysis of quantitative proteomics data. The main problems arising at each step of the analysis are discussed conceptually and instead of presenting all methods developed by the community, MaxQuant [Cox and Mann, 2008] and Perseus [Tyanova et al., 2016] often serve as examples on how these problems could be addressed. The manuscript was written with Jürgen Cox and Pavel Sinitcyn. Furthermore, I contributed figures and to the writing of the peptide identification and the statistical downstream analysis sections.

Pavel Sinitcyn, Jan Daniel Rudolph, and Jürgen Cox. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science*, 1(1):annurev–biodatasci–080917–013516, July 2018a. ISSN 2574-3414. doi: 10.1146/annurev-biodatasci-080917-013516

Annual Review of Biomedical Data Science

Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data

Pavel Sinitcyn,* Jan Daniel Rudolph,* and Jürgen Cox

Computational Systems Biochemistry Research Group, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany; email: cox@biochem.mpg.de

Annu. Rev. Biomed. Data Sci. 2018. 1:207–34

First published as a Review in Advance on May 4, 2018

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-080917-013516

*These authors contributed equally to this article

Keywords

computational proteomics, mass spectrometry, posttranslational modifications, multiomics data analysis, multivariate analysis, network analysis

Abstract

Computational proteomics is the data science concerned with the identification and quantification of proteins from high-throughput data and the biological interpretation of their concentration changes, posttranslational modifications, interactions, and subcellular localizations. Today, these data most often originate from mass spectrometry–based shotgun proteomics experiments. In this review, we survey computational methods for the analysis of such proteomics data, focusing on the explanation of the key concepts. Starting with mass spectrometric feature detection, we then cover methods for the identification of peptides. Subsequently, protein inference and the control of false discovery rates are highly important topics covered. We then discuss methods for the quantification of peptides and proteins. A section on downstream data analysis covers exploratory statistics, network analysis, machine learning, and multiomics data integration. Finally, we discuss current developments and provide an outlook on what the near future of computational proteomics might bear.

INTRODUCTION

Proteins perform nearly all the work in a cell and are the key players in the structure, function, and regulation of cells, tissues, and organs. Collectively they form the proteome (1), a highly dynamic and diverse molecular omics space comprising interactions among proteins and other types of biomolecules. The proteome can be studied comprehensively with mass spectrometry (MS)-based technologies (2–4). Thousands of proteins and posttranslational modifications (PTMs) can be studied quantitatively over a multitude of samples in complex experimental designs. Describing all applications of proteomics is beyond the scope of this review, but among its applications are diverse topics such as cancer immunotherapy (5) and the evolution of extinct species (6).

Computational MS-based proteomics can be roughly subdivided into two main areas: (*a*) the identification and quantification of peptides, proteins, and PTMs and (*b*) downstream analysis, aiming at the biological interpretation of the quantitative results obtained in area *a*. This review follows this subdivision. Computational proteomics is a highly multidisciplinary endeavor attracting scientists from many fields and incorporates other disciplines like statistics, machine learning, efficient scientific programming, and network and time series analysis. Furthermore, the integration of proteomics data with other biological high-throughput data is increasingly gaining importance.

Peptide-based shotgun proteomics, also called bottom-up proteomics (7), needs to be distinguished from top-down proteomics (8–10), in which whole proteins are studied in the mass spectrometer. Data analysis tools and approaches exist for top-down methods (11–13) in which feature deconvolution plays an important part. In targeted proteomics (14–17) (**Figure 1**), a set of key peptides from a target list, which is informative for a set of proteins or PTMs of interest, is quantitatively monitored over many samples using dedicated software (18). Data-independent acquisition (19), as exemplified by the SWATH-MS method, comes with its own computational challenges for which solutions are provided in the literature (20–23). Imaging MS (24) is also a

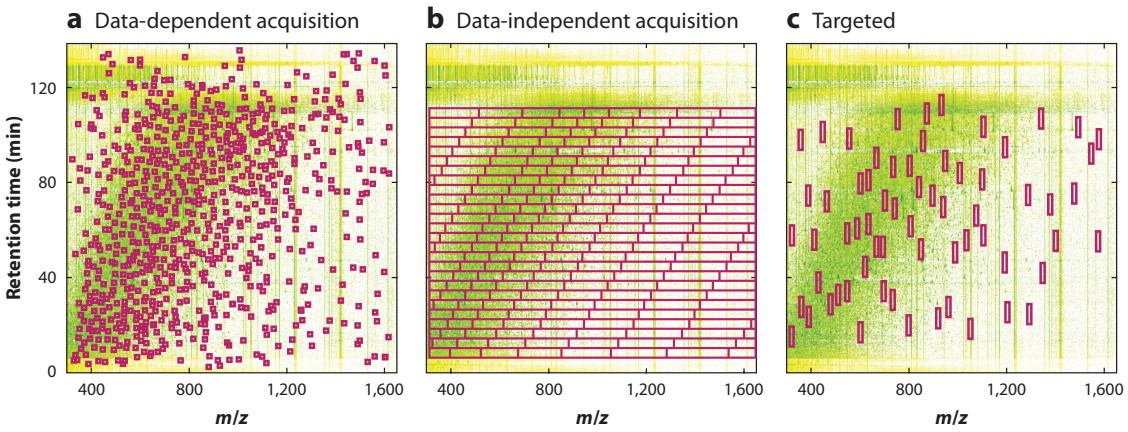


Figure 1

Main formats of mass spectrometry (MS)-based proteomics. Peptide-based bottom-up proteomics is most often done in the data-dependent acquisition mode (*a*). MS2 (second-stage MS) scans are triggered depending on the MS1 (first-stage MS) data features seen in real time. Typically, at a given retention time, the *n* most intense peptide features are selected for fragmentation, dynamically excluding masses that have just been previously selected. In data-independent acquisition (*b*), a set of constant mass ranges, which do not depend on the peptides being analyzed, is isolated for fragmentation. In targeted proteomics (*c*), a list of peptides is targeted based on a list of mass and retention time ranges corresponding to peptides of interest, which are particularly informative of a set of proteins or posttranslational modifications that are the focus of the investigation.

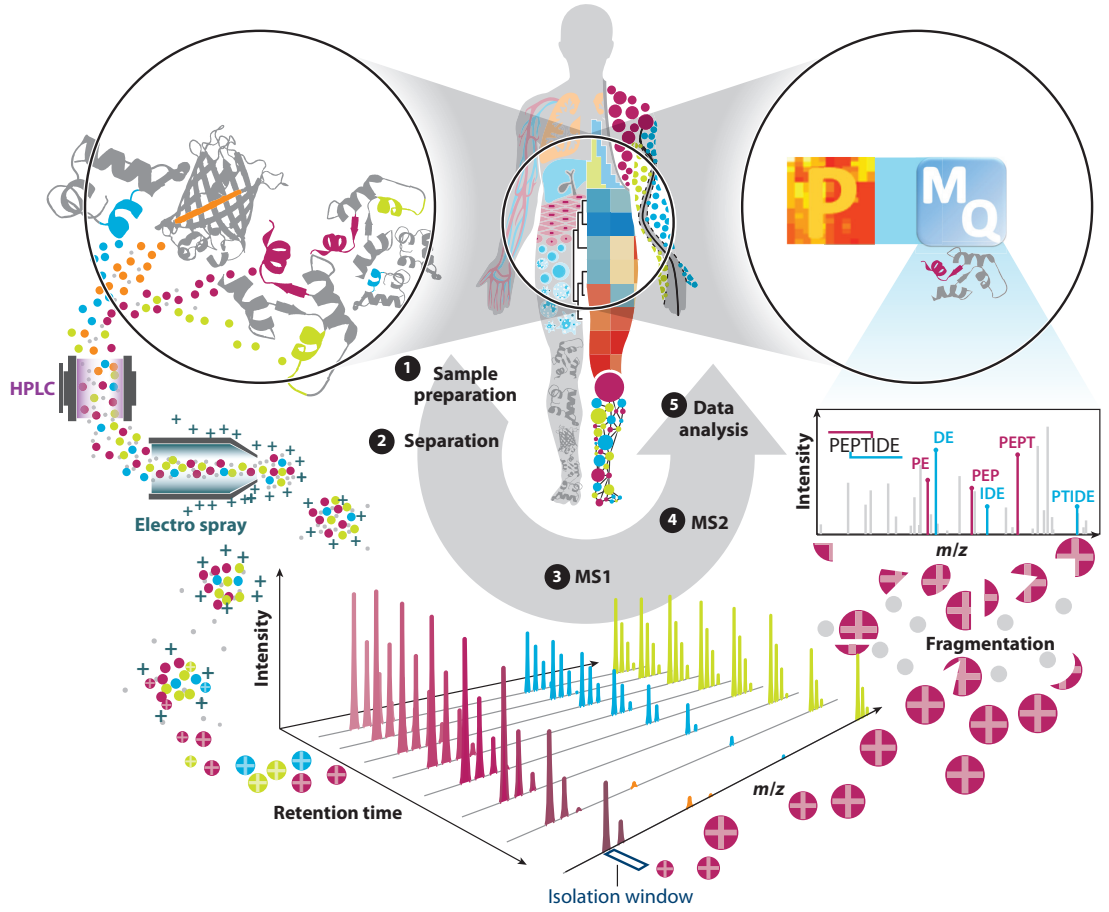


Figure 2

Bottom-up shotgun proteomics workflow. (❶) Proteins are extracted from a sample of interest. Enrichment of organelles or affinity purification may be performed. Proteins are digested to peptides that are optionally enriched for modifications. (❷) After HPLC separation, peptides are ionized (181, 182) and (❸) injected into a high-resolution mass spectrometer (e.g., 183, 184). MS1 spectra containing peptide isotope patterns are recorded in a cycle with a timescale of about one second. (❹) Peptide precursors are selected for fragmentation and fragment (MS2) spectra are recorded. (❺) Both MS1 and MS2 spectra are written to disk, typically resulting in several gigabytes of data per LC-MS run, and then analyzed by computational proteomics software. Abbreviations: HPLC, high-performance liquid chromatography; LC, liquid chromatography; MS, mass spectrometry; MS1, first-stage MS; MS2, second-stage MS.

fruitful area of research that will not be covered here. This review focuses on data-dependent bottom-up or shotgun proteomics (**Figure 2**), which currently is the format most frequently used in proteomics.

It is not the aim of this review to present an exhaustive list of all available software tools. Instead, we focus on explaining concepts and key applications. In several places, we use the MaxQuant (25–27) and Perseus (28) software as concrete examples for the implementation of certain concepts. Alternative software platforms developed in academia (29–31) or offered by mass spectrometer vendors can provide similar functionality. We propose that robustness, ease of use, parallelizability,

and automation of all computational aspects are the key factors to consider in the selection of software tools.

Proteomics research is supported by community tools such as repositories, databases, and annotation sources (32). There are public repositories for the storage and dissemination of MS-based proteomics data (33–39), and submission of raw data is highly recommended for every proteomics publication (34). Protein and peptide sequences are essential for the interpretation of proteomics data. For this purpose, UniProt (universal protein resource) (40) is a comprehensive, high quality, and freely accessible resource of protein sequences and functional information. Since most amino acid sequence identifications can be put into the context of coding nucleic acid sequences—exceptions prove the rule (41)—genome-centric sequence repositories like Ensembl (42) are of high importance as well. Data sharing and dissemination of publicly available proteomics data are facilitated by dedicated software tools for the reanalysis of community data (43, 44).

This review consists of two main parts, the first dealing with the data analysis steps performed on the spectral data itself, going up to the identification and quantification of peptides, proteins, and PTMs. This part is organized in a problem-centric way, where in each subsection, a particular challenge in the MS workflow is described. The second part is about the downstream data analysis. Here, the sections are organized by methodologies rather than application areas, which is a more approachable organization scheme, since the number of different applications is enormous, while the methodologies overlap. The downstream analysis of proteomics data is still an art, and there is not always only one correct way to arrive at biologically meaningful conclusions. Hence, we give a comprehensive overview of the available methods that can be used along the way.

IDENTIFICATION AND QUANTIFICATION OF PEPTIDES, PROTEINS, AND POSTTRANSLATIONAL MODIFICATIONS

Liquid Chromatography-Mass Spectrometry Features

Since the early days of MS, the detection of peaks in a mass spectrum, corresponding to molecular features, played a central role (45). Nowadays, the mass resolution is sufficiently high in general that the isotope pattern of peptides is resolvable (**Figure 3*a***). On the molecular level, a single peak corresponds to an isotopic species with fixed elemental composition and several nucleons. In case of ultrahigh mass resolution, the isotopic fine structure of peptides in the low-mass range can be resolved (46) (**Figure 3*a***), resulting in increased information about the atomic constituents of the peptide. While obtaining isotopic resolution is standard nowadays for peptides, the same is still technically challenging for whole proteins in top-down proteomics. For instance, for each charge state of an antibody, usually only an envelope is detected, while the isotopic peaks remain unresolved.

In proteomics, the mass spectrometer is typically coupled on-line to additional continuous separation dimensions like liquid chromatography (LC) (47) or ion mobility separation (48). MS features can therefore be viewed as higher-dimensional objects. In case of LC-MS, peaks become three-dimensional (3D) objects in the m/z–retention time–intensity space (**Figure 3*b***). Using ion mobility adds another dimension, turning features into 4D objects. Technically, due to its dimensionality, the problem of MS feature detection is equivalent to general-purpose 2D image feature detection or voxel assembly to 3D volume elements (49), respectively. However, since MS data often have additional regularities that can be exploited, the problem is often simpler than generic object recognition. Simplifying assumptions specific to mass spectrometer types should be exploited to apply faster algorithms to the multidimensional feature detection problem. (Readers are referred to the supplement of Reference 25.)

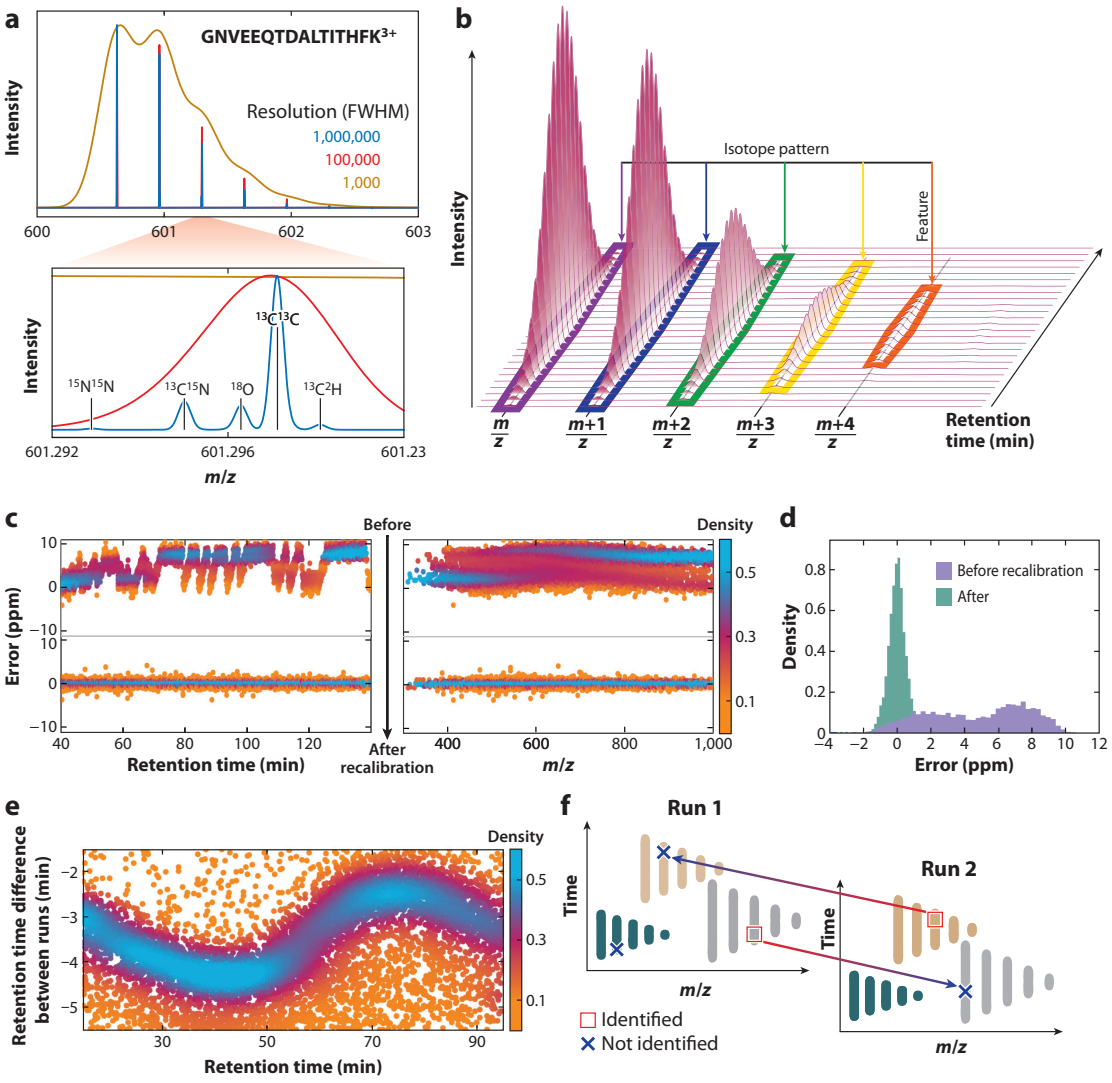


Figure 3

MS1 feature–based computational tasks in a proteomics workflow. (*a*) Theoretical spectrum of an MS1 feature measured in three different resolutions. The lowest resolution (1,000 FWHM) does not resolve the isotope pattern. The ultrahigh resolution (1,000,000) reveals the natural isotopic fine structure. (*b*) A three-dimensional isotope pattern in *m/z*–retention time–intensity space. (*c*) Peptide mass errors as a function of retention time and peptide *m/z* before and after nonlinear recalibration. Clearly, nonlinear systematic errors were present and were then removed by recalibration. (*d*) Mass error distribution before and after recalibration. A large increase in mass accuracy was achieved through nonlinear recalibration. (*e*) Retention time alignment curve between two LC-MS runs. (*f*) Matching between runs. Peptide identities are transferred between LC-MS runs from MS2-identified MS1 features to nonidentified MS1 features in other similar LC-MS runs based on accurate mass and retention time. Abbreviations: FWHM, full width at half maximum; LC, liquid chromatography; MS, mass spectrometry; MS1, first-stage MS; MS2, second-stage MS; ppm, parts per million.

Once features corresponding to isotopic peaks are detected, they are assembled to isotope patterns, effectively deisotoping the spectrum. Different models exist (50–52), one of them being the Averagine model (50), which can be used to explore spectral properties, since nearly all peptides with a given approximate molecular mass have a similar elemental composition. In the model, it is assumed that a peptide is made up of the average number of the 20 amino acids according to their natural occurrence. The model then predicts the mass differences between isotopic peaks in an isotope pattern, as well as their relative heights. This approach is usually sufficient when dealing with data with unresolved isotopic fine structure. When the isotopic fine structure is resolved, one will have to employ the true atomic compositions of the peptide candidates to utilize this information. In the approaches using higher-dimensional features, the exact coelution of isotopic peaks can also be utilized to increase the specificity of assignment of isotope patterns. While in most cases, the spectral information is not sufficient to determine the elemental composition, one will obtain the charge state and a highly precise estimate of the monoisotopic mass from the information contained in the higher-dimensional features.

One can find labeling *n*-plexes of isotope patterns in the MS1 (first-stage MS) data prior to peptide identification, similar to how features are assembled to isotope patterns. This applies to nonradioactive differential isotopic sample labeling techniques (53, 54) like SILAC (stable isotope labeling by amino acids in cell culture) (55) or dimethyl labeling (56, 57). Analogous to the deisotoping step, specific mass differences between the isotope patterns participating in a labeling *n*-plex are expected. This is not the case for ^{15}N labeling (58, 59) in which all nitrogen atoms are completely exchanged with the stable heavy isotope. Isotope patterns belonging to an *n*-plex are usually coeluting, depending on the type of labeling, which can be exploited in the assembly of *n*-plexes.

While mass measurements from modern high-resolution mass spectrometers, in combination with the aforementioned higher-dimensional feature detection, can achieve very-high-mass precision, this does not automatically translate into high-mass accuracies, due to the presence of systematic measurement errors. In **Figure 3*c***, the peptide mass error prior to mass recalibration is displayed as functions of *m/z* and of retention time. Systematic errors are typically nonlinear and depend on multiple variables. In addition to *m/z* and retention time, the mass error can depend on signal intensity and ion mobility index, if applicable. Nonlinear recalibration on multidimensional parameters is difficult when it must rely on only a few calibration points, as is usually the case if dedicated spike-in molecules are used. Hence, it is typically better in complex samples to use the peptides from the sample itself as calibration points for multivariate recalibration, which is achieved in MaxQuant by a two-level peptide identification strategy (25, 60, 61). The mass accuracy increases by large factors resulting from the applications of these nonlinear recalibration curves obtained in this way (**Figure 3*d***).

Similar to the mass accuracy, the consistency of the retention times of peptide features can also be increased by recalibration. Due to often unavoidable irreproducibility in chromatography, retention times are usually not comparable between LC-MS runs, thereby limiting identification-transfer and quantification between runs. Nonlinear shifts by several minutes are common. Hence, algorithmic approaches were developed to align retention times between multiple runs (**Figure 3*e***). Typically, these retention time corrections need to be nonlinear (62). In MaxQuant, this is achieved with a sample similarity–derived guide tree, which avoids the need for singling out one LC-MS run as the master run (63) that all the other runs are aligned to. Ion mobilities can be aligned between LC-MS runs with similar methods as retention times.

Once masses, retention times, and ion mobilities are recalibrated, one can transfer identifications between related LC-MS runs from peptide features identified by fragmentation to unidentified peptide features by having same mass, charge, retention time, and ion mobility (64)

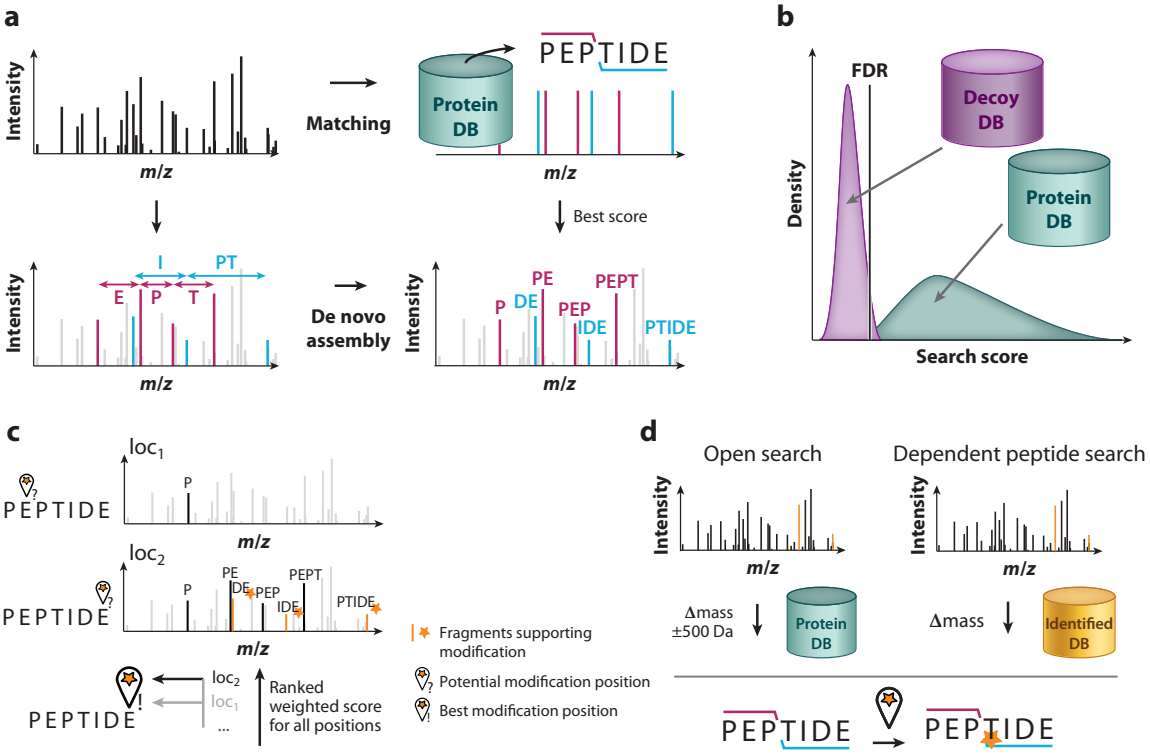
(**Figure 3*f***). Following this strategy, the quantification profiles across many samples become more complete, which partially removes the stochastic behavior of the data-dependent acquisition in bottom-up proteomics. Determining and controlling false discovery rates (FDRs) for these kind of matching approaches is challenging and the subject of current research. However, if samples are similar, error rates caused by matching are in acceptably low ranges.

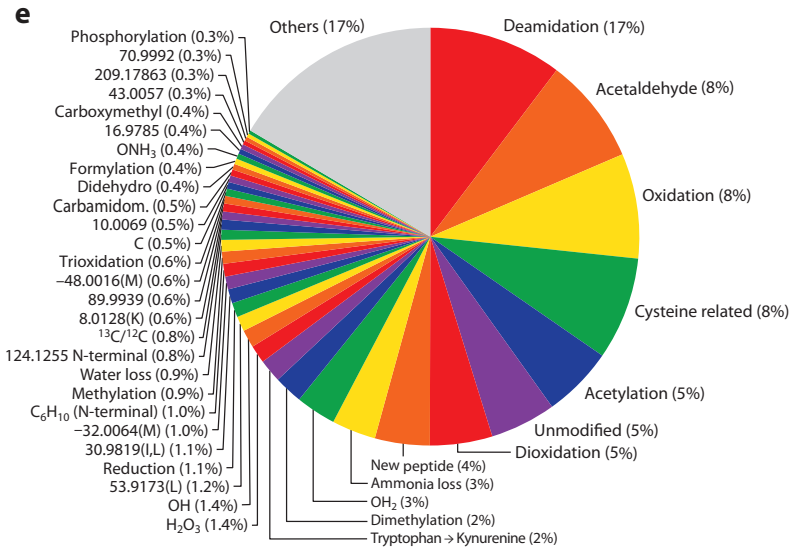
Peptide Identification

Peptide identification tools analyze the fragmentation spectra obtained by the mass spectrometer with the aim of determining the sequence of the peptide. In the most popular approach, database search engines (65–69) utilize a target database of theoretical fragmentation for identification (**Figure 4*a***). The database is generated from all protein sequences that are known or thought to be produced according to the instructions in the genome of an organism. The protein sequences are digested in silico into peptides according to a cleavage rule mirroring the protease used in the experiment (e.g., trypsin, which cleaves after the occurrence of lysine or arginine in the protein sequence). For each of these in silico peptides, the list of expected fragment masses is calculated based on the backbone bond breakages expected for the fragmentation technique used in the experiment. For a given measured fragmentation spectrum, the search engine calculates a match score against all theoretical fragmentation spectra within a specified peptide mass tolerance. The highest-scoring peptide spectrum match (PSM) is taken as a candidate for the identity of the peptide. Since the highest-scoring PSM might still be a false positive, most workflows control the FDR using a target–decoy approach (70) (**Figure 4*b***). In this approach, fragmentation spectra are searched not only against the target database, but also against a decoy database, which is designed to produce false-positive PSMs. Comparing the score distributions of target and decoy PSMs, posterior error probabilities can be calculated and FDRs can be controlled. One procedure to generate decoy sequences is to reverse the target sequences, providing peptides that do not occur in nature.

Additional peptide features besides the search engine score, such as the length of the peptide and the number of missed cleavages, help distinguish true identifications from false positives, leading to more high-confidence identifications. In MaxQuant, the posterior error probability, which is the probability of a PSM being wrongly identified, is conditional on the score and additional peptide properties (25). Other tools such as PeptideProphet (71, 72) and Percolator (73) use linear discriminant analysis or support vector machines (SVMs) with the same aim. Machine learning was used to predict intensity patterns in fragmentation spectra in order to support database scoring and further improve identification (74), but it failed to improve upon the state of the art. In contrast, the application of deep learning to de novo peptide identification did yield improvements (75).

De novo peptide sequencing (**Figure 4*a***) is another technique for identifying peptides from fragmentation spectra. The peptide is identified using only information from the input spectrum and the characteristics of the fragmentation method. Mass differences between certain peak pairs correspond to amino acid masses, which are interpreted as consecutive ions in one of the expected fragment series, for example, y or b ions for collision-induced dissociation. If these mass differences can be continued to a whole series from N- to C-termini, the peptide is identified without reference to a sequence database. An incomplete de novo amino acid series is called a sequence tag and might be completed on either of the termini with a sum of amino acid masses and PTMs. The many existing tools for de novo peptide identification explore different algorithmic approaches, some allowing for de novo sequencing errors and homology searches (76–79). An interesting approach is a hybrid between database search and de novo sequencing (80); it requires only a little de novo information and hence inherits high sensitivity from the database search approach.





(Caption appears on following page)

Figure 4 (*Figure appears on preceding page*)

Overview of peptide identification methods. (*a*) In the peptide database (DB) search engine approach, measured second-stage mass spectrometry (MS2) spectra are scored against a list of theoretical spectra from an in silico digest of protein sequences. De novo peptide identification allows reading the peptide sequence partially or completely out of the MS2 spectrum. (*b*) In the target–decoy approach, true and decoy protein sequences are offered to estimate the false discovery rate (FDR). (*c*) Determining the localization probability for a posttranslational modification on a peptide. (*d*) Open search and dependent peptide search are methods for detecting modifications in an unbiased way. Modifications still must be localized after open search. (*e*) Modifications found in a typical dependent peptide search. Data from Reference 185 were used.

For a peptide that has been identified as having a certain sequence and carrying one or more modifications, the positions of these modifications on the sequence might not be localizable with complete certainty. Hence, a score needs to be calculated that quantifies for each potentially modifiable amino acid in the peptide sequence the certainty of localization at a given locus (**Figure 4*c***). For instance, a peptide might contain several potentially phosphorylated serine, threonine, and tyrosine residues, but from the peptide mass it is known that it is phosphorylated only once. Then one needs to determine which of the sites are phosphorylated and use the spectral evidence to derive each site's probability that it is the one bearing the modification (81–85). The most important spectral features for the calculation of localization probabilities are the site-determining ions, which are fragments that are matched with one hypothetical localization but not with the other. The exact way the localization score is calculated varies between different methods. In MaxQuant, the localization probability is calculated as a weighted average of exponential Andromeda scores over all combinations of phosphorylation configurations (86).

The identification of modified amino acids, either as PTMs such as phosphorylation or as modifications introduced during sample preparation, is usually done by adding these as variable modifications into the database search. While this strategy is highly sensitive, all modifications have to be specified beforehand. The number of modifications that can be specified is limited due to the combinatorial explosion of modified peptides species, leading to a large increase in database size. There are two approaches overcoming these limitations: open search (87) and dependent peptide search (88) (**Figure 4*d***). The open search approach does not extend the sequence database but instead widens the precursor mass tolerance window for the MS1 precursor peptide molecule to, for example, ±500 Da, while keeping the fragment mass tolerance low (87). Therefore, a modified peptide with a mass within the tolerance window can still be matched to the correct unmodified database sequence despite ∼50% of fragment ions being shifted by the modification. The high number of candidate matches makes the open search computationally demanding, but recent approaches make use of fragment ion indexing to speed up the search significantly (89). The dependent peptide search, also implemented in MaxQuant, is a generic approach to retrospectively identify unassigned MS2 (second-stage MS) scans; it relies on the assumption that the sample contains not only the modified dependent peptide, but also its unmodified base peptide counterpart (88). Using any search algorithm will yield identifications, as well as unassigned MS2 spectra. The search now queries all unassigned spectra against all identified spectra, while simultaneously localizing the modification. The mass difference between the peptides is the putative mass of the modification, which is used to generate a shifted ion series for each position in the peptide. The highest-scoring match will therefore determine the sequence of the peptide, as well as the mass and locus of the modification. **Figure 4*e*** shows the most frequent modifications found by dependent peptide search in a typical data set.

There are a number of special topics in peptide identification, starting with dipeptides resulting from cross-linked proteins (90, 91), which have the challenge of a vastly increased search space due to pairing of peptides, for which several popular software packages are available (92–97). In proteogenomics searches (98), peptides are identified based on customized protein sequence

databases generated from genomic or transcriptomic information. Search spaces for proteogenomics searches are typically larger than in conventional searches since they often involve three- or six-frame translations of genomic sequences. Furthermore, these search spaces are heterogeneous, since the sequence content ranges from clearly existing, manually validated protein sequences to in silico–translated genomic regions without any prior evidence for their expression. Hence, extra measures need to be taken in the identification process to account for this heterogeneity. Proteomics of species without sequenced genome requires tools to integrate incomplete sequencing data with homologous sequence data from closely related species (99).

Protein Inference and False Discovery Rate

Protein inference, that is, the assembly of peptides into a list of proteins, is a crucial step in a computational proteomics workflow, since usually the peptides are only technical aids to study proteins. (Readers are referred to Reference 100 for a review.) The relationship between peptides and proteins is many-to-many, since upon digestion a protein gives rise to many peptides, but a peptide can also originate from more than one protein. Furthermore, based on the identified peptides, proteins that share common sequences might not be distinguishable from each other. Hence, a redundancy grouping of protein sequences is necessary.

Peptides that are unique to a protein are more desirable than nonunique ones. On average, longer peptides are more likely to be unique, and hence, more informative. As an order of magnitude estimate, we calculate how often a random peptide of a given length would occur in the human proteome, assuming it is randomly composed out of the 20 amino acids and has the same size as the latest human UniProt release 2017_09, which contains 93,588 protein sequences comprising 37,118,756 amino acids in total. Peptides of length 5 should occur on average 12 times in the proteome, meaning that their information content is nearly worthless. Peptides of length 6 should occur on average 0.6 times, making them only just potentially useful, but many of them can still be expected to be nonunique. In this model, only peptides of length 7 or longer are on average expected to be informative and useful. Although other factors like tryptic peptides and paralog relationships between genes realistically should be considered, the conclusions hold true of real data.

Many tools and algorithms for the protein assembly have been described in the literature. The most frequently applied ones can be roughly subdivided into parsimonious and statistical models. Parsimonious models (25, 101–104) apply Occam's razor principle (105) to the protein inference problem by finding a set of proteins that is as small as possible to explain the observed peptides. Usually, fast greedy heuristics are used to find such a protein set. Statistical models (106, 107) can assemble large amounts of weak peptide identifications to infer the existence of a protein. However, for both types of models, it is worth considering a threshold on peptide identification quality, for example, 1% FDR for PSMs. High-quality peptide identifications allow for solid conclusions about the properties of the identified proteins, while weakly identified peptides can compromise protein quantification accuracy. Ideally, the output of the protein inference step is a list of protein groups. Each protein group contains a set of proteins that cannot be distinguished from each other based on the observed peptides. Either the proteins in a protein group have equal sets of identified peptides or the peptide set of one protein is a proper subset of that of another protein, in which case, based on the peptide identifications, there is no evidence for the existence of the latter protein, assuming that the former protein is in the sample.

The phenomenon of error expansion from peptide to protein identification in large data sets is well known in the field (106, 108). Even if the FDR is thoroughly controlled at the PSM level, if no additional measures are taken, the FDR on protein level can become arbitrarily large. Hence, it

is highly important to use workflows that control FDR on the protein level (25, 106, 108, 109) to limit the number of proteins falsely claimed to be present in the sample, particularly if the number of identified proteins is a relevant outcome of the study.

Quantification

Proteomics becomes more powerful when done quantitatively, as compared to only browsing through lists of identified proteins. Many responses to stimuli on the level of proteins are not switching the expression of a protein on and off completely, but manifest themselves as changes in cellular concentrations that might be small, yet important. Quantitative proteomics approaches can be subdivided into absolute and relative quantification methods. In absolute quantification, one wants to determine copy numbers or concentrations of proteins within a sample, while in relative quantification, a quantitative ratio or relative change of protein concentrations between samples is desired. Both absolute and relative quantification can be done either with the aid of labels or label-free.

Figure 5 shows an overview of relative quantification methods. In label-free quantification, the samples being compared are biochemically processed separately. The distinction between metabolic and chemical labeling is not important from a computational perspective. Instead, the main distinction is between MS1-level labeling, in which the peptide signals corresponding to the multiple samples are compared and form multiplexed isotope patterns in the MS1 spectra, and MS2-level or isobaric labeling, in which the multiplexed signals appear in the fragmentation spectra. Hence, computational methods for relative quantification should be distinguished between label-free, MS1-level labeling, and MS2-level labeling.

In label-free quantification, one faces particular challenges with normalization intensities between LC-MS runs and the compatibility of quantification with prefractionation. In MaxQuant, the MaxLFQ algorithm (110) is implemented for relative label-free quantification. It uses signal intensities of MS1 peptide features as input, optionally including the ones identified by matching between runs, and produces as output relative protein abundance profiles over multiple samples. MaxLFQ accounts for any peptide or protein prefractionation of the samples by applying a sophisticated intensity normalization procedure to the feature intensities of each LC-MS run. A protein intensity profile is constructed that best fits protein ratios determined in all pairwise comparisons between samples. In each of these pairwise comparisons, only peptides that occur in both samples are used, which makes the relative comparison very precise. Hence, MaxLFQ is more accurate than merely summing up all peptide intensities belonging to a protein. By using a sample-similarity network for the intensity normalization step, the algorithm scales well to large data sets and can quantify hundreds of samples against each other.

Stable isotope labeling with sample multiplexing appearing on the level of MS1 spectra (55–57, 111, 112) promises to be more accurate than label-free quantification since the coelution of features in the same LC-MS run can be exploited. The ratio calculation can be performed along the elution profile separately in each MS1 scan and separately for each isotopic peak. This results in many estimates of the ratio, which can be summarized by taking the median. This robust ratio estimate is less sensitive to contamination by other coeluting peptides. In this way, the ratios between MS1-label channels are calculated in a more precise way, as compared to the label-free approach, where feature intensities are calculated separately before their ratio is taken. During MS1-label n-plex assembly, the isotope patterns of parts of the n-plex might be missing, leading to an incomplete quantitative profile. Proper MS1 isotope patterns might be missing for peptides arising from low-abundant proteins. In MaxQuant, the requantification algorithm tries to find traces of these isotope patterns close to the noise level.

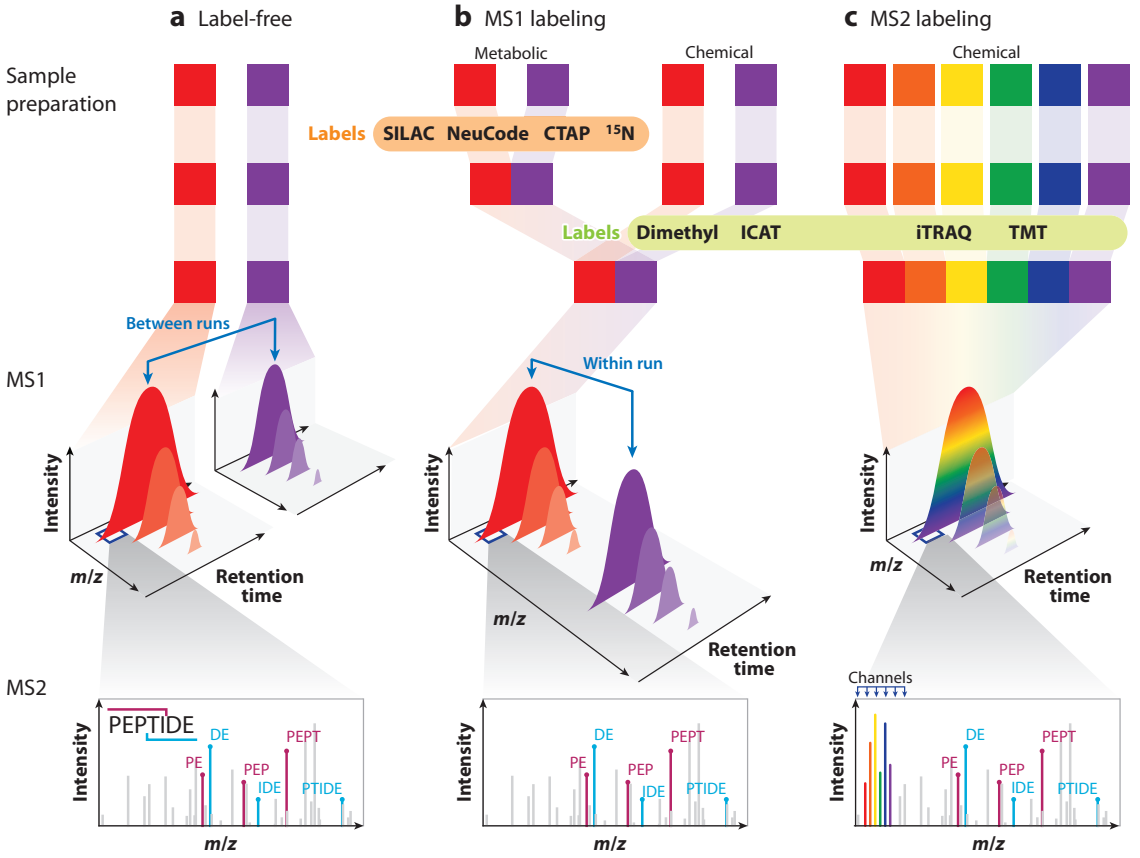


Figure 5

Overview of relative quantification methods. Relative quantification of samples (*colored squares*) can be done in a label-free, metabolic, or chemical labeling approach. For computational approaches, the distinction between MS1 labeling (*b*) and MS2 (isobaric) labeling (*c*) is more crucial. In the label-free approach (*a*), the quantification is done for each peptide feature between extracted ion chromatograms in different LC-MS runs. In MS1 label–based quantification (e.g., SILAC, dimethyl, NeuCode), multiple samples will appear as differentially labeled isotope patterns in the MS1 spectra. For isobaric labeling (e.g., iTRAQ, TMT), the quantification signals appear as reporter ions in the low-mass range of the MS2 spectra. Abbreviations: CTAP, cell type–specific labeling using amino acid precursors; ICAT, isotope-coded affinity tags; iTRAQ, isobaric tags for relative and absolute quantification; LC, liquid chromatography; MS, mass spectrometry, MS1, first-stage MS; MS2, second-stage MS; SILAC, stable isotope labeling with amino acids in cell culture; TMT, tandem mass tags.

One can use one labeling channel as a common standard, as is done in Super-SILAC (113), which allows quantifying unlabeled samples with the added accuracy of labeling by using ratios of ratios to compare samples with each other. Computationally, these hybrid samples are analyzed like MS1-labeled samples in the feature detection, but the downstream analysis proceeds nearly as if they were label-free samples.

In isobaric labeling (114–116), peptides in different samples are labeled with different molecules per sample that have the same mass but that eject different reporter ions upon fragmentation. The biggest advantage of isobaric labeling is its multiplexing capacity. Up to 11 samples can be measured simultaneously with the currently available tandem mass tag reagents. The downside is

that the presence of coeluting peptides in the isolation window for fragmentation leads to ratio compression (117). To be precise, cofragmentation makes ratios wrong in arbitrary and individual ways. However, since it is often a valid assumption that most of the proteins are not changing between samples, the cofragmented peptides are likely to have 1:1 ratios, thus compressing the ratios of changing proteins. There are several experimental strategies to reduce or remove the cofragmentation problem, such as gas-phase purification (118), MultiNotch MS3 (119), and use of complementary ions (120). There are several computational methods that reduce ratio compression. Reporter ions of low intensity are prone to carry more noise and be more affected by cofragmentation signals. Hence, peptides with higher reporter ion intensities should be given higher weights when calculating protein intensities. Another approach is to calculate the fraction of precursor signal divided by the total MS1 signal observed in the isolation window (121, 122), which can be used for filtering peptides used for quantification. To some extent, this quantity can also be used to correct for ratio compression (123).

Approximate measures of absolute protein abundances can be obtained with simple computational prescriptions like the iBAQ or Top3 methods (124, 125). The problem that peptides of a protein have vastly different flyability (a term used to cover the relative efficiencies of ionization, transfer, and detection), making them not directly comparable for quantification, is solved by averaging over many peptides or selecting the most intense ones, which enriches for high flyability. For eukaryotic cells, one can add an absolute scale to these readouts with the proteomic ruler approach (126), which uses the signal of histones, assuming that it is proportional to the amount of DNA in the sample.

The quantification of peptides and PTMs differs from protein quantification in that only a single or few features can be used for quantification, while on the protein level, accuracy is achieved by accumulating quantitative information over many peptides. Hence, the variability of PTM quantification data and the number of missing values is usually higher than it is for proteins. For combined PTM-enriched and proteome data, computational methods exist for calculating occupancies (86, 127), which are the percentages of proteins modified at a given PTM site.

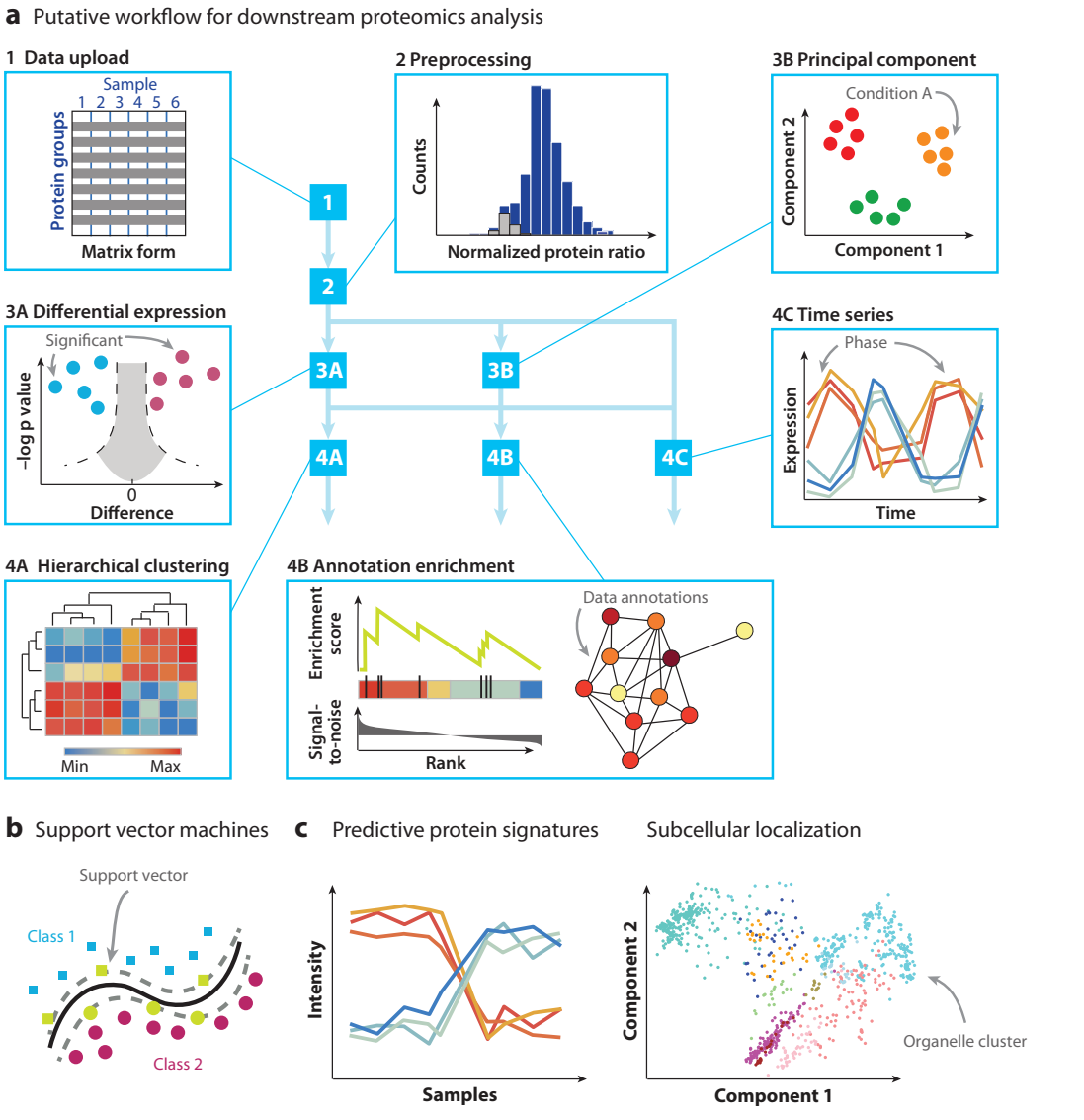
DOWNSTREAM DATA ANALYSIS

Exploratory Statistics

Once proteins have been identified and quantified over many samples, one obtains a matrix with proteins (or protein groups) as rows, samples as columns, and protein abundances or abundance ratios in the matrix cells. Usually, the interpretation of this quantitative protein or PTM data and the translation into significant biological or biomedical findings are the most important and labor-intensive parts of a study. The Perseus platform (28) was developed to support the domain expert in this data exploration. It is workflow based, modular, and extensible through a plugin infrastructure.

There are some preparatory steps preceding most analyses, such as normalization of intensities or ratios, data filtering, and potentially missing-value imputation (**Figure 6*a***). A common task in discovery proteomics is to identify proteins of biological interest and distinguish them from the rest of the proteome. Statistical models are popular tools for identifying differentially expressed proteins. Clustering methods, such as hierarchical clustering, are often used for finding expression patterns of groups of proteins and for their visualization in a heat map. Principal component analysis (PCA) is an alternative method of visualizing the main effects in the data and the relatedness between samples. It also provides information on proteins responsible for a separation of sample groups through the so-called loadings.

The statistical tests *t*-test and ANOVA (analysis of variance, which is the generalization of the *t*-test to more than two groups) are the basic versions of a series of statistical models that test for significant changes between sample groups (128, 129). In more complex experimental designs, one might want to test for the effects of two factors simultaneously (e.g., gender and treatment), in which case two-way ANOVA can be used. ANOVA can be generalized to any number *n* of factors, resulting in *n*-way ANOVA. After retrieving a list of significant proteins from ANOVA, a post hoc test can be applied to pinpoint the sample groups within the experimental design that were changing. If samples are related and independency assumptions are violated, so-called



(Caption appears on following page)

Figure 6 (*Figure appears on preceding page*)

Downstream analysis overview. (*a*) Putative workflow for downstream proteomics analysis. After data upload (*Step 1*) and preprocessing (*Step 2*), common analyses include differential expression (*Step 3A*), principal component analysis (*Step 3B*), hierarchical clustering (*Step 4A*), annotation enrichment (*Step 4B*) and time series analysis (*Step 4C*). Data preprocessing (*Step 2*) may involve several steps including data normalization and visual inspection of distributions of protein quantification values in histograms. Differential expression analysis (*Step 3A*) reveals those proteins that are significantly changing their concentrations between two or more conditions. Principal component analysis (*Step 3B*) highlights main trends in the data such as a separation between cellular conditions, as shown in the example. Hierarchical clustering (*Step 4A*) is often done in conjunction with heat map visualization of expression changes and reveals characteristic patterns relating groups of samples to clusters of proteins. Results are often validated using annotation enrichment analysis (*Step 4B*). Time series analysis (*Step 4C*) can distinguish between characteristic temporal patterns such as phases of peaking protein concentrations in a periodic process, as shown in the example. Adapted from Reference 28. (*b*) Support vector machines are a powerful machine learning tool for classification. From training data they learn decision rules that can distinguish between classes of samples based on their protein expression profiles. The decision rule is indicated here by a separating line between the two classes. Support vectors are those samples that contribute most to defining the separating line. Adapted from Reference 28. (*c*) Applications for machine learning in proteomics include finding predictive protein signatures and predicting the subcellular localization of proteins. The colored clusters represent proteins that are localized in same organelles. Data from Reference 147 were used.

repeated measures ANOVA is a valid method of data analysis. For all of the methods above, it is crucial to control false positives due to multiple hypothesis testing, since many tests are done simultaneously. If only a moderate *p*-value cutoff is applied to define significant proteins, the number of false positives will be inflated (130). Benjamini-Hochberg FDR control (131) or permutation-based FDR estimates (132) are efficient methods to deal with this problem.

When an interesting group of proteins has been identified, for instance, by statistical testing, clustering, or PCA, enrichment analysis can be performed to find biological processes, complexes, or pathways common to these proteins. Fisher's exact test checks for contingency between group membership and the property of interest. It clarifies what is common to the cluster-member proteins and might indicate the functional role of the cluster. For this purpose, annotation sources like gene ontology (133), pathway memberships (134), or curated protein complexes (135) are needed.

Biological processes under study often exhibit temporal changes, with proteins following an expected pattern, for instance, as periodic changes in the cell cycle or circadian rhythm. Other studies involve measuring a response to dose changes of stimuli. In these situations, methods can be applied that detect concentration changes following a given model, such as periodic changes with a given periodicity. For this case of periodic temporal changes, the analysis will assign an amplitude of change and a peaking time to each protein (136).

Posttranslational Modifications

Quantitative PTM data can be represented as a matrix resembling proteome-expression data, but with modified peptides or modification sites on the identified proteins as rows. Therefore, PTM studies can be analyzed with methods similar to those used for protein expression. For instance, after suitable normalization and filtering, hierarchical clustering or PCA can be applied to determine dominant patterns of phosphorylation changes (86). As previously discussed, one needs to be aware of the higher variance of PTM-level data compared to protein-level data. This requires a higher number of replicates compared to protein-level data to achieve the same statistical power.

There are several public resources for obtaining PTM specific annotations. UniProt (40) provides comprehensive information on local protein properties at the PTM site or in its vicinity. Specialized databases, such as PhosphoSitePlus (137), Signor (138), and Phospho.ELM (139), cover mostly phosphorylation events. They include functional annotations, as well as kinase–substrate interactions. This information can be used for enrichment analysis to gain information about the processes involved in writing, reading, and erasing the studied PTMs. One can also analyze PTMs in the context of signaling networks, as discussed below.

Machine Learning

Machine learning has several applications in the downstream analysis of proteomics data (**Figure 6*b*,*c***). A very prominent one is the classification of patient-derived samples based on their protein expression patterns (140–142). For artificial intelligence–based diagnosis, a supervised learning algorithm is first trained on samples derived from patient cohorts for which a certain property is known, for instance, the cancer subtype. The trained algorithm is then used to diagnose novel samples, that is, to predict the same property for samples where the property is not known. The same supervised learning approach can be combined with feature selection algorithms to derive predictive protein signatures. Each signature contains proteins that show a distinct expression pattern and can be used for sample classification. Multivariate feature selection methods can take the interdependence of proteins acting within networks into account and can find patterns for which the discriminatory power is not apparent in the expression profiles of single proteins. This makes machine learning–based feature selection a powerful alternative to ANOVA-like methods to determine protein signatures, where a *p*-value is calculated for only one protein at a time, independently from all the other proteins.

Machine learning approaches are most easily validated using cross-validation (143), which provides a measure of how well the prediction performance of a classification or regression model will generalize to independent data not used for model training. Cross-validation helps avoid the notorious problem of model overfitting and can be used to monitor prediction errors when extracting optimal protein sets from the output of feature selection algorithms. SVMs (144) often perform particularly well in classification or regression of samples in omics spaces. This is not surprising, since for most technologies, including proteomics, the number of features (biomolecules) is typically much larger than the number of samples. SVMs were created to perform well in spaces with exactly these properties. Deep learning (145, 146) is gaining traction in proteomics (75) and will likely find more applications in the future.

Machine learning has also been successfully applied to the prediction of subcellular localization with the dynamic organellar maps method (147, 148), which allows global mapping of protein translocation events. First, one generates a database of marker proteins with known localization and absolute copy number information and characteristic fractionation profiles. Then, using SVMs, a model is built for the prediction of cellular localization. This method has dynamic capabilities to capture translocation events upon a stimulation. This enables a widely applicable proteome-wide analysis of cellular protein movements without requiring process-specific reagents.

Network Biology

MS-based proteomics provides researchers with diverse tools for the study of biological networks (149). Enrichment protocols interrogate the interaction partners of a bait protein and provide the basis for the assembly of large-scale protein–protein interaction (PPI) networks (**Figure 7*a***). Affinity enrichment/purification coupled to LC-MS is routinely used to quantify hundreds of physical interaction partners. Since relying only on identification of proteins in the pull-down leads to many false positives, it is crucial to distinguish background binders from significantly enriched bona fide interactors. Statistical tests, such as the two-sample *t*-test, can identify true interactors but require a control to compare against. This control sample either can be a dedicated experiment lacking the bait protein or can be assembled from other orthogonal experiments within the same study (150, 151). Due to its quantitative nature, this approach can probe not only steady-state interactions, but also dynamic rewiring upon stimulation by internal or external stimuli. If intensity-based quantification is used, the missing values problem for enriched samples can be overcome by imputation. Alternative methods

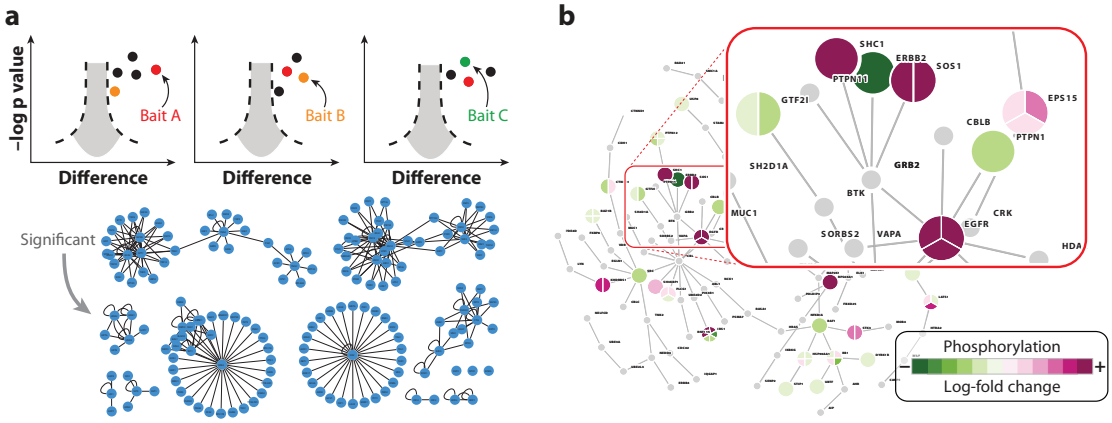


Figure 7

Network analysis. (*a*) Protein–protein interaction networks can be constructed by applying statistical testing to a series of pull-down experiments with different bait proteins. The resulting network of proteins with significant enrichment to any of the bait proteins can be visualized in tools such as Cytoscape. Adapted from Reference 28. (*b*) Signaling pathway reconstructed from phosphoproteomics data derived from MCF7 cells after epidermal growth factor stimulation (160). The pie charts in the network visualize the measured phosphorylation changes on each of the proteins. Proteins with unknown phosphorylation states are colored gray.

relying on spectral counting directly accommodate for the absence or presence of a protein in a sample (152). Both approaches have been used to construct large-scale PPI networks (151, 153).

Cells often achieve signal transduction through PTMs, which are enzymatically written, read, and erased. The interpretation of PTMs in the context of these signaling networks is therefore natural. PTM specific networks, such as kinase–substrate interactions, can be obtained from curated databases, such as PhosphoSitePlus (137). To increase coverage, kinase–substrate relationships can also be predicted by machine learning and PPI network analysis (154). Logic models obtained from, for example, the Signor database (138) can provide a mechanistic interpretation of phosphoproteomic data, indicating active kinases, as well as functional phosphorylation sites. Several computational methods predict kinase activities from kinase–substrate interactions and phosphoproteomics data. For a recent review and benchmark, readers are referred to References 155 and 156. Kinase–substrate enrichment analysis (157) uses parametric tests to compare the changes of the substrates of one kinase to all other substrates. Cluster evaluation (158) clusters phosphorylation sites based on time series data, from which enrichments of kinase–substrate annotations are calculated. Inference of kinase activities from phosphoproteomics (159) uses machine learning to estimate the strength of kinase–substrate interactions, as well as kinase activities. Phosphoproteomic dissection using networks (PHOTON) (160) is a method using general PPI networks for interpreting phosphorylation data within their signaling context. PHOTON identifies proteins that significantly contribute to signaling and uses these proteins to reconstruct the most plausible signaling pathway from the PPI network (**Figure 7*b***).

For general-purpose network analysis, Cytoscape (161) has emerged as the de facto standard. Through its plugin infrastructure, it provides a wealth of analyses and visualizations, often integrating expression-omics technologies with interaction networks. Cytoscape reads networks from various standard formats and can extend them with interactions and pathways from various databases. Tools such as BiNGO (162) can identify significantly enriched gene ontology

terms in these networks. Large-scale networks can be clustered into modules, either by topology (MCODE; see Reference 163) or by differential expression (jActiveModules; see Reference 164). Alternatively, network reconstruction tools, such as ANAT (165), identify a subset of interactions connecting, for example, differentially expressed proteins to their signaling stimulus.

Multiomics Data Analysis

Analyzing data from two omics technologies applied to the same samples becomes straightforward if there is a near one-to-one match between the biomolecules measured in each of the two omics spaces. For instance, when comparing the proteome and the transcriptome, the one-to-one correspondence between transcript and protein sequences holds true with only little deviations due to, for example, translation errors and postprocessing of the protein sequence. Thus, the molecular correspondence is sufficiently valid to conceptually work with matching rows between the two omics matrices. The problem reduces to mapping transcript to protein identifiers and to dealing with the different depth in distinguishable splice variants, for which algorithmic solutions exist (28). A similar molecular correspondence can be applied to the genome–proteome spaces for correlating local genomic properties such as DNA copy number (166) or loss of heterozygosity to protein expression if proteins matching to the same gene model are grouped together. Also, ribosomal profiling data (167) can be brought into molecular correspondence with proteomics data.

Once a correspondence between omics spaces has been established, one can perform pointwise comparisons, as is done in the scatterplots in **Figure 8*a***, in which protein abundances, messenger RNA levels, and ribosomal profiling data are compared. Individual outliers in each of these plots may hint at interesting biology. However, it is difficult to assign significance to individual data points. Hence, researchers developed 2D annotation enrichment (168; **Figure 8*b***) to answer the question, Which classes of gene products show concordant and which show discordant behavior between the different levels of gene expression? While transcriptional regulation is a dominant factor in expression control, there are many known examples of posttranscriptional regulation like microRNA-controlled inhibition of transcripts (169) and directed protein degradation (170), which are detectable by this method.

Further examples of simultaneous multivariate analysis in two matched omics spaces are joint time series analysis, which is exemplified in **Figure 8*c*** for circadian transcriptomics, and proteomics data (136). Here, it was possible to derive time lags between peaks in transcript and protein abundances as a proxy for the time lag between transcription and translation for individual cycling transcripts and their associated proteins. Additionally, joint transcriptomics–proteomics PCA performed on the same data (**Figure 8*d***) indicates global similarities in transcript and protein concentrations, but with a time delay.

When the input is time-resolved data for transcriptome and proteome, protein expression control analysis (PECA) (171, 172) computes the probability of regulation changes between adjacent time intervals. PECA quantitatively dissects protein expression variation into the contributions of mRNA and protein synthesis–degradation rate ratios.

Unlike in the previous examples, when combining proteomics with metabolomics, there is not a one-to-one correspondence between molecules. In biochemical pathways, proteins are associated with reactions between metabolites as catalysts. The required mapping of biomolecules is facilitated by the consensus human metabolic reconstruction Recon 2.2 (173), which has a high potential for integrating and analyzing diverse data types. Recon 2.2 facilitates the integration of proteomics data with an updated curation of relationships between genes, proteins, and reactions.

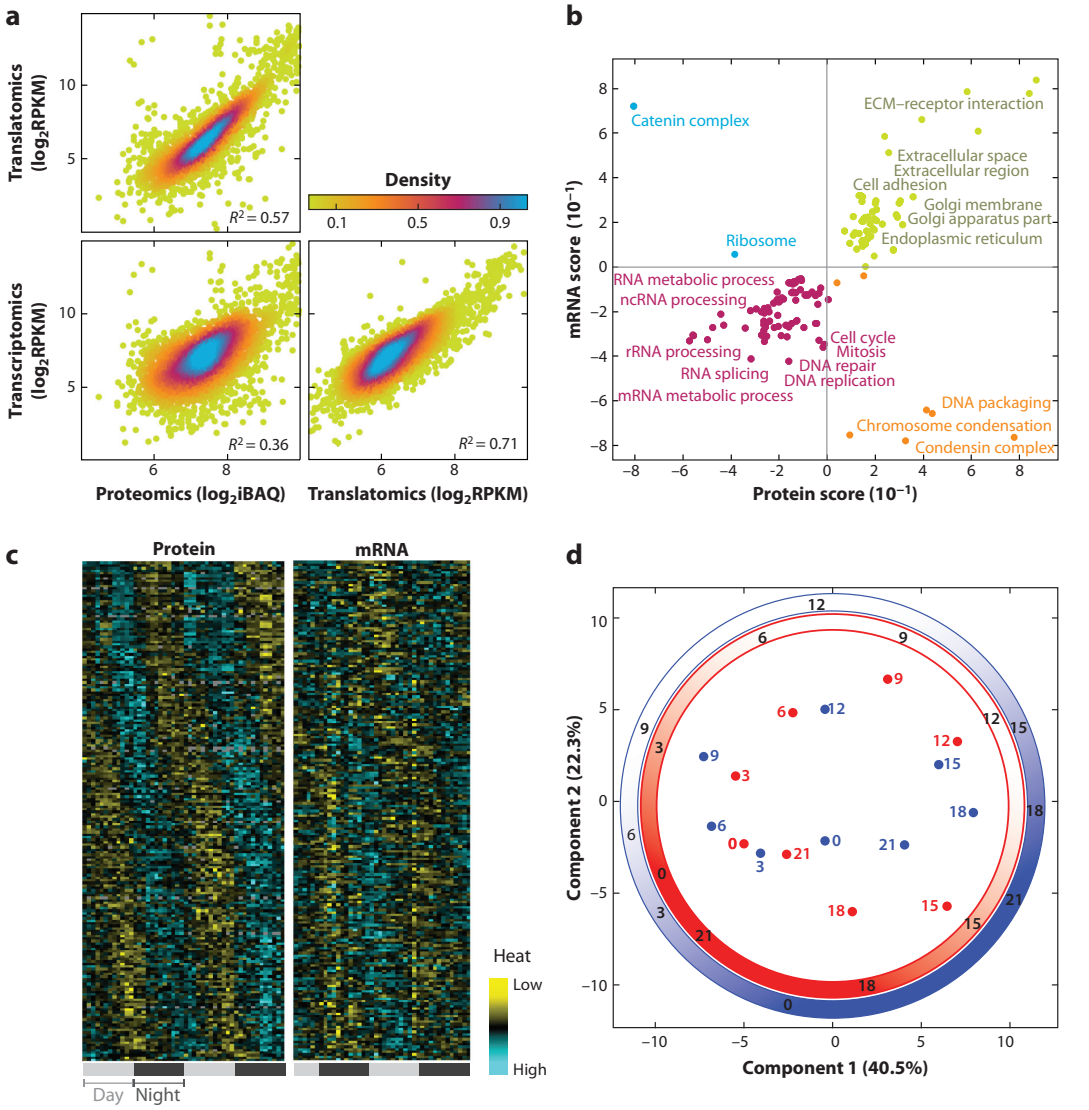


Figure 8

Cross-omics data analysis. (*a*) Comparison of protein abundances, ribosomal profiling data, and mRNA expression. Proteins are quantified with the iBAQ method (124), while RPKM (186) was used for the other two data types. Adapted from Reference 28. (*b*) Output of the two-dimensional enrichment analysis applied to protein and mRNA abundances. Adapted from Reference 28. (*c*) Side-by-side heat maps for daily rhythmic proteins and transcripts showing a cycling pattern. In the rows, samples are ordered by time of extraction, and in the columns, proteins are ordered by time of their peak concentration. Adapted from Reference 136. (*d*) Principal component analysis performed jointly on transcriptomics data (*red*) and proteomics data (*blue*) of two phases of circadian mouse liver data. Labels next to data points denote time in hours. Both transcriptomics and proteomics data points arrange in a periodic time series pattern in the first two principal components. Adapted from Reference 136. Abbreviations: ECM, extracellular matrix; iBAQ, intensity-based absolute quantification; mRNA, messenger RNA; ncRNA, noncoding RNA; RPKM, reads per kilobase per million mapped reads; rRNA, ribosomal RNA.

DISCUSSION AND OUTLOOK

Computational proteomics has matured substantially and is keeping up well with the massive amounts of data produced by modern mass spectrometers. Platforms for identification and quantification of proteins can analyze the data in a reliable and automated way. Therefore, attention is increasingly being shifted to the downstream part of the data analysis, in which the quantification results are interpreted, hypotheses are tested, and novel biological and biomedical knowledge is gained. We anticipate that future developments of computational proteomics tools will be particularly active in these areas, including network biology and cross-omics data analysis. In previous work (28), we made the case for enabling the end users—the researchers from fundamental biology, drug discovery, and medical sciences—to perform large parts of the data analysis themselves, and this is increasingly happening.

Single-cell DNA and RNA sequencing (174) have shed new light onto the heterogeneity and diversity of biological processes behind the cellular averages that are typically monitored in many omics technologies. According to reports in the literature (175), single-cell proteomics is just around the corner and will likely bear many new discoveries. Once it is scalable and sufficiently deep in terms of proteome coverage, it might help define a highly resolved atlas of all cell types and cell states in the human body (176). Certainly, novel computational tools will have to be developed for the particular challenges of single-cell proteomics data, which will likely have unique challenges in terms of normalization and handling of missing data.

There is still a large gap between the generation of large-scale proteomics data and the modeling of signaling pathways and biochemical reactions. The curated knowledge of PTMs currently available in public resources (134, 177) is still limited and needs to be expanded to support more comprehensive analyses. New tools are emerging to reconstruct signaling pathways and translate them into logic models (178). Hopefully, the path from large-scale time series data to kinetic modeling (179, 180) will become more accessible for many interdisciplinary researchers, leading to an improved mechanistic understanding of the biological processes under investigation based on large-scale data.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement 686547 and from the FP7 grant agreement GA ERC-2012-SyG_318987–ToPAG.

LITERATURE CITED

1. James P. 1997. Protein identification in the post-genome era: the rapid rise of proteomics. *Q. Rev. Biophys.* 30(4):279–331
2. Cox J, Mann M. 2011. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* 80:273–99
3. Altelaar AF, Munoz J, Heck AJ. 2013. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* 14(1):35–48

4. Aebersold R, Mann M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature* 537(7620):347–55
5. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, et al. 2016. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* 7:13404
6. Welker F, Collins MJ, Thomas JA, Wadsley M, Brace S, et al. 2015. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature* 522(7554):81–84
7. Wolters DA, Washburn MP, Yates JR. 2001. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 73(23):5683–90
8. Fornelli L, Durbin KR, Fellers RT, Early BP, Greer JB, et al. 2017. Advancing top-down analysis of the human proteome using a benchtop quadrupole-orbitrap mass spectrometer. *J. Proteome Res.* 16(2):609–18
9. Toby TK, Fornelli L, Kelleher NL. 2016. Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem.* 9:499–519
10. Chait BT. 2006. Mass spectrometry: bottom-up or top-down? *Science* 314(5196):65–66
11. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, et al. 2007. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* 35:W701–6
12. Kou Q, Xun L, Liu X. 2016. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* 32(22):3495–97
13. Park J, Piehowski PD, Wilkins C, Zhou M, Mendoza J, et al. 2017. Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* 14(9):909–14
14. Gillette MA, Carr SA. 2013. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nat. Methods* 10(1):28–34
15. Liebler DC, Zimmerman LJ. 2013. Targeted quantitation of proteins by mass spectrometry. *Biochemistry* 52(22):3797–3806
16. Picotti P, Aebersold R. 2012. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* 9(6):555–66
17. Ebhardt HA, Root A, Sander C, Aebersold R. 2015. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* 15(18):9193–208
18. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, et al. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26(7):966–68
19. Doerr A. 2014. DIA mass spectrometry. *Nat. Methods* 12(1):35–35
20. Rosenberger G, Bludau I, Schmitt U, Heusel M, Hunter CL, et al. 2017. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* 14(9):921–27
21. Bruderer R, Bernhardt OM, Gandhi T, Xuan Y, Sondermann J, et al. 2017. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteom.* 16(12):2296–309
22. Bilbao A, Varesio E, Luban J, Strambio-De-Castillia C, Hopfgartner G, et al. 2015. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* 15(5–6):964–80
23. Tsou C-C, Avtonomov D, Larsen B, Tucholska M, Choi H, et al. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* 12(3):258–64
24. McDonnell LA, Heeren RMA. 2007. Imaging mass spectrometry. *Mass Spectrom. Rev.* 262007:606–43
25. Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26(12):1367–72
26. Tyanova S, Temu T, Cox J. 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 11(12):2301–19
27. Tyanova S, Temu T, Carlson A, Sinitcyn P, Mann M, Cox J. 2015. Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics* 15(8):1453–56
28. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, et al. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* 13(9):731–40
29. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, et al. 2016. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Meth.* 13(9):741–48

30. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, et al. 2010. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10(6):1150–59
31. McIlwain S, Tamura K, Kertesz-Farkas A, Grant CE, Diament B, et al. 2014. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* 13(10):4488–91
32. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaíno JA. 2015. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 15(5–6):930–50
33. Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, et al. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 44(D1):D447–56
34. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, et al. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32(3):223–26
35. Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, et al. 2014. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteom.* 13(10):2765–75
36. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509(7502):582–87
37. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. 2014. A draft map of the human proteome. *Nature* 509(7502):575–81
38. Schaab C, Geiger T, Stoehr G, Cox J, Mann M. 2012. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell. Proteom.* 11(3):M111.014068
39. Desiere F. 2006. The PeptideAtlas project. *Nucleic Acids Res.* 34(90001):D655–58
40. UniProt Consort. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45(D1):D158–69
41. Mohimani H, Yang YL, Liu WT, Hsieh PW, Dorrestein PC, Pevzner PA. 2011. Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* 11(18):3642–50
42. Yates A, Akanni W, Amode MR, Barrell D, Billis K, et al. 2016. Ensembl 2016. *Nucleic Acids Res.* 44(D1):D710–16
43. Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. 2011. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 11(5):996–99
44. Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, et al. 2015. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* 33(1):22–24
45. Zhang J, Gonzalez E, Hestilow T, Haskins W, Huang Y. 2009. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr. Genom.* 10(6):388–401
46. Miladinović SM, Kozhinov AN, Gorshkov MV, Tsybin YO. 2012. On the utility of isotopic fine structure mass spectrometry in protein identification. *Anal. Chem.* 84(9):4042–51
47. Snyder LR, Kirkland JJ, Dolan JW. 2010. *Introduction to Modern Liquid Chromatography*. Hoboken, NJ: Wiley
48. Kanu AB, Dwivedi P, Tam M, Matz L, Hill HH. 2008. Ion mobility–mass spectrometry. *J. Mass Spectrom.* 43(1):1–22
49. Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. 2006. Cluster-based analysis of FMRI data. *Neuroimage* 33(2):599–608
50. Senko MW, Beu SC, McLafferty FW. 1995. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 6(4):229–33
51. Rockwood AL, Van Orden SL, Smith RD. 1996. Ultrahigh resolution isotope distribution calculations. *Rapid Commun. Mass Spectrom.* 10(1):54–59
52. Horn DM, Zubarev RA, McLafferty FW. 2000. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 11(4):320–32
53. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. 1999. Accurate quantitation of protein expression and site-specific phosphorylation. *PNAS* 96(12):6591–96
54. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. 2007. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 389(4):1017–31
55. Ong SE, Mann M. 2007. Stable isotope labeling by amino acids in cell culture for quantitative proteomics. *Methods Mol. Biol.* 359:37–52

56. Hsu JL, Huang SY, Chow NH, Chen SH. 2003. Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* 75(24):6843–52
57. Boersema PJ, Aye TT, van Veen TA, Heck AJ, Mohammed S. 2008. Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates. *Proteomics* 8(22):4624–32
58. Engelsberger WR, Erban A, Kopka J, Schulze WX. 2006. Metabolic labeling of plant cell cultures with $K^{15}NO_3$ as a tool for quantitative analysis of proteins and metabolites. *Plant Methods* 2(3):14
59. Ippel JH, Pouvreau L, Kroef T, Gruppen H, Versteeg G, et al. 2004. In vivo uniform ^{15}N-isotope labelling of plants: using the greenhouse for structural proteomics. *Proteomics* 4(1):226–34
60. Cox J, Michalski A, Mann M. 2011. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* 22(8):1373–80
61. Cox J, Mann M. 2009. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *J. Am. Soc. Mass Spectrom.* 20(8):1477–85
62. Podwojski K, Fritsch A, Chamrad DC, Paul W, Sitek B, et al. 2009. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* 25(6):758–64
63. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, et al. 2007. *SuperHirn*—a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 7(19):3470–80
64. Pasa-Tolic L, Masselon C, Barry RC, Shen Y, Smith RD. 2004. Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques* 37(4):621–36
65. Eng JK, McCormack AL, Yates JR. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5(11):976–89
66. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18):3551–67
67. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. 2004. Open mass spectrometry search algorithm. *J. Proteome Res.* 3(5):958–64
68. Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20(9):1466–67
69. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 10(4):1794–1805
70. Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4(3):207–14
71. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74(20):5383–92
72. Choi H, Nesvizhskii AI. 2008. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* 7(1):254–65
73. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* 4(11):923–25
74. Degroeve S, Martens L, Jurisica I. 2013. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* 29(24):3199–203
75. Tran NH, Zhang X, Xin L, Shan B, Li M. 2017. De novo peptide sequencing by deep learning. *PNAS* 114(31):8247–52
76. Taylor JA, Johnson RS. 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 11(9):1067–75
77. Ma B, Zhang K, Hendrie C, Liang C, Li M, et al. 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 17(20):2337–42
78. Ma B, Johnson R. 2012. *De novo* sequencing and homology searching. *Mol. Cell. Proteom.* 11(2):O111.014902
79. Han Y, Ma B, Zhang K. 2004. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *Proc. Comput. Syst. Bioinform. Conf., Stanford, Calif., 16–19 Aug.*, pp. 206–15. New York: IEEE
80. Bern M, Cai Y, Goldberg D. 2007. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* 79(4):1393–1400

81. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* 24(10):1285–92
82. Bailey CM, Sweet SMM, Cunningham DL, Zeller M, Heath JK, Cooper HJ. 2009. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. *J. Proteome Res.* 8(4):1965–71
83. Lemeer S, Kunold E, Klaeger S, Raabe M, Towers MW, et al. 2012. Phosphorylation site localization in peptides by MALDI MS/MS and the Mascot Delta Score. *Anal. Bioanal. Chem.* 402(1):249–60
84. Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, et al. 2011. Confident phosphorylation site localization using the Mascot Delta Score. *Mol. Cell. Proteom.* 10(2):M110.003830
85. Taus T, Köcher T, Pichler P, Paschke C, Schmidt A, et al. 2011. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* 10(12):5354–62
86. Sharma K, D'Souza RC, Tyanova S, Schaab C, Wisniewski JR, et al. 2014. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* 8(5):1583–94
87. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, et al. 2015. A mass-tolerant database search—supplementary. *Nat. Biotechnol.* 33(7):743–49
88. Savitski MM. 2006. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteom.* 5(5):935–48
89. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. 2017. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* 14(5):513–20
90. Sinz A. 2006. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. *Mass Spectrom Rev.* 25(4):663–82
91. Singh P, Panchaud A, Goodlett DR. 2010. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Anal. Chem.* 82(7):2636–42
92. Hoopmann MR, Zelter A, Johnson RS, Riffle M, MacCoss MJ, et al. 2015. Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res.* 14(5):2190–98
93. Götze M, Pettelkau J, Schaks S, Bosse K, Ihling CH, et al. 2012. StavroX—a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* 23(1):76–87
94. Liu F, Lössl P, Scheltema R, Viner R, Heck AJR. 2017. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* 8:15473
95. Yang B, Wu YJ, Zhu M, Fan SB, Lin J, et al. 2012. Identification of cross-linked peptides from complex samples. *Nat. Methods* 9(9):904–6
96. Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, et al. 2010. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell Proteom.* 9(8):1634–49
97. Chen ZA, Fischer L, Cox J, Rappsilber J. 2016. Quantitative cross-linking/mass spectrometry using isotope-labeled cross-linkers and MaxQuant. *Mol. Cell Proteom.* 15:2769–78
98. Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 11(11):1114–25
99. Temu T, Mann M, Räschle M, Cox J. 2016. Homology-driven assembly of NOn-redundant protEin sequence sets (NOmESS) for mass spectrometry. *Bioinformatics* 32(9):1417–19
100. Huang T, Wang J, Yu W, He Z. 2012. Protein inference: a review. *Brief. Bioinform.* 13(5):586–614
101. Yang X, Dondeti V, Dezube R, Maynard DM, Geer LY, et al. 2004. DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.* 3(5):1002–8
102. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, et al. 2009. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* 8(8):3872–81
103. Slotta DJ, McFarland MA, Markey SP. 2010. MassSieve: panning MS/MS peptide data for proteins. *Proteomics* 10(16):3035–39
104. Alves P, Arnold RJ, Novotny MV, Radivojac P, Reilly JP, Tang H. 2007. Advancement in protein inference from shotgun proteomics using peptide detectability. *Proc. Pac. Symp. Biocomput., Maui, Hawaii, 3–7 Jan.*, pp. 409–20. **http://psb.stanford.edu/psb-online/proceedings/psb07/alves.pdf**
105. Sober E. 2017. *Ockham's Razors: A User's Manual*. Cambridge, UK: Cambridge Univ. Press
106. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75(17):4646–58

107. Serang O, MacCoss MJ, Noble WS. 2010. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* 9(10):5346–57
108. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, et al. 2009. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell Proteom.* 8(11):2405–17
109. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M. 2015. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell Proteom.* 14:2394–404
110. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteom.* 13(9):2513–26
111. Gauthier NP, Soufi B, Walkowicz WE, Pedicord VA, Mavrakis KJ, et al. 2013. Cell-selective labeling using amino acid precursors for proteomic studies of multicellular environments. *Nat. Methods* 10(8):768–73
112. Merrill AE, Hebert AS, MacGilvray ME, Rose CM, Bailey DJ, et al. 2014. NeuCode labels for relative protein quantification. *Mol. Cell Proteom.* 13(9):2503–12
113. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. 2010. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* 7(5):383–85
114. Thompson A, Schäfer JJ, Kuhn K, Kienle S, Schwarz J, et al. 2003. Tandem mass tags: a novel quantificaiton strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75(8):1895–1904
115. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. 2004. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteom.* 3(12):1154–69
116. Rauniyar N, Yates JR. 2014. Isobaric labeling-based relative quantification in shotgun proteomics. *J. Proteome Res.* 13(12):5293–303
117. Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC. 2009. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly." *J. Proteome Res.* 8(11):5347–55
118. Wenger CD, Lee MV, Hebert AS, McAlister GC, Phanstiel DH, et al. 2011. Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* 8(11):933–35
119. McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, et al. 2014. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* 86(14):7150–58
120. Wühr M, Haas W, McAlister GC, Peshkin L, Rad R, et al. 2012. Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* 84(21):9214–21
121. Savitski MM, Fischer F, Mathieson T, Sweetman G, Lang M, Bantscheff M. 2010. Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. *J. Am. Soc. Mass Spectrom.* 21(10):1668–79
122. Michalski A, Cox J, 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. *J. Proteome Res.* 10(4):1785–93
123. Savitski MM, Mathieson T, Zinn N, Sweetman G, Doce C, et al. 2013. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* 12(8):3586–98
124. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. 2011. Global quantification of mammalian gene expression control. *Nature* 473(7347):337–42
125. Silva JC. 2005. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteom.* 5(1):144–56
126. Wisniewski JR, Hein MY, Cox J, Mann M. 2014. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol. Cell Proteom.* 13(12):3497–506
127. Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, et al. 2010. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* 3(104):ra3
128. Krzywinski M, Altman N. 2013. Points of significance: significance, P values and t-tests. *Nat. Methods* 10:1041–42

129. Krzywinski M, Altman N. 2014. Points of significance: Analysis of variance and blocking. *Nat. Methods* 11(7):699–700
130. Noble WS. 2009. How does multiple testing correction work? *Nat. Biotechnol.* 27(12):1135–37
131. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300
132. Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98(9):5116–21
133. Gene Ontol. Consort. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43:D1049–56
134. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, et al. 2016. The reactome pathway knowledgebase. *Nucleic Acids Res.* 44(D1):D481–87
135. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, et al. 2009. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* 38(Suppl.1):D646–50
136. Robles MS, Cox J, 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(1):e1004047
137. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* 43:D512–20
138. Perfetto L, Briganti L, Calderone A, Perpetuini AC, Iannuccelli M, et al. 2016. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.* 44(D1):D548–54
139. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, et al. 2011. Phospho.ELM: A database of phosphorylation sites—update 2011. *Nucleic Acids Res.* 39(Suppl. 1):D261–67
140. Deeb SJ, Tyanova S, Hummel M, Schmidt-Supprian M, Cox J, Mann M. 2015. Machine learning based classification of diffuse large B-cell lymphoma patients by their protein expression profiles. *Mol. Cell Proteom.* 14(11):2947–60
141. Iglesias-Gato D, Wikstrom P, Tyanova S, Lavallee C, Thysell E, et al. 2015. The proteome of primary prostate cancer. *Eur. Urol.* 69(5):942–52
142. Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M, Geiger T. 2016. Proteomic maps of breast cancer subtypes. *Nat. Commun.* 7:10259
143. Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. Int. Jt. Conf. Artif. Intell., 14th, Montr., Can., 20–25 Aug.*, pp. 1137–43. San Francisco: Morgan Kaufmann
144. Vapnik VN. 1995. *The Nature of Statistical Learning Theory*. New York: Springer
145. Schmidhuber J. 2015. Deep learning in neural networks: an overview. *Neural Netw.* 61:85–117
146. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44
147. Itzhak DN, Tyanova S, Cox J, Borner GHH. 2016. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* 5:e16950
148. Itzhak DN, Davies C, Tyanova S, Mishra A, Williamson J, et al. 2017. A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell Rep.* 20(11):2706–18
149. Bensimon A, Heck AJR, Aebersold R. 2012. Mass spectrometry–based proteomics and network biology. *Annu. Rev. Biochem.* 81:379–405
150. Keilhauer EC, Hein MY, Mann M. 2015. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteom.* 14(1):120–35
151. Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, et al. 2015. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163(3):712–23
152. Sowa ME, Bennett EJ, Gygi SP, Harper JW. 2009. Defining the human deubiquitinating enzyme interaction landscape. *Cell* 138(2):389–403
153. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, et al. 2015. The BioPlex network: a systematic exploration of the human interactome. *Cell* 162(2):425–40
154. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, et al. 2008. NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* 36(Suppl. 1):D695–99
155. Dermit M, Dokal A, Cutillas PR. 2017. Approaches to identify kinase dependencies in cancer signalling networks. *FEBS Lett.* 591(17):2577–92

156. Hernandez-Armenta C, Ochoa D, Gonçalves E, Saez-Rodriguez J, Beltrao P. 2017. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* 33(12):1845–51
157. Casado P, Rodriguez-Prados J-C, Cosulich SC, Guichard S, Vanhaesebroeck B, et al. 2013. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* 6(268):rs6
158. Yang P, Zheng X, Jayaswal V, Hu G, Yang JYH, Jothi R. 2015. Knowledge-based analysis for detecting key signaling events from time-series phosphoproteomics data. *PLOS Comput. Biol.* 11(8):e1004403
159. Mischnik M, Sacco F, Cox J, Schneider HC, Schäfer M, et al. 2015. IKAP: A heuristic framework for inference of kinase activities from phosphoproteomics data. *Bioinformatics* 32(3):424–31
160. Rudolph JD, de Graauw M, van de Water B, Geiger T, Sharan R. 2016. Elucidation of signaling pathways from large-scale phosphoproteomic data using protein interaction networks. *Cell Syst.* 3(6):585–93
161. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13(11):2498–2504
162. Maere S, Heymans K, Kuiper M. 2005. *BiNGO*: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21(16):3448–49
163. Bader GD, Hogue CW. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* 4:2
164. Ideker T, Ozier O, Schwikowski B, Siegel AF. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl. 1):S233–40
165. Yosef N, Zalckvar E, Rubinstein AD, Homilius M, Atias N, et al. 2011. ANAT: a tool for constructing and analyzing functional protein networks. *Sci. Signal.* 4(196):pl1
166. Geiger T, Cox J, Mann M. 2010. Proteomic changes resulting from gene copy number variations in cancer cells. *PLOS Genet.* 6(9):e1001090
167. Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* 15(3):205–13
168. Cox J, Mann M. 2012. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinform.* 13(Suppl. 1):S12
169. He L, Hannon GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* 5(7):522–31
170. Hochstrasser M. 1996. Ubiquitin-dependent protein degradation. *Annu. Rev. Genet.* 30:405–39
171. Teo G, Vogel C, Ghosh D, Kim S, Choi H. 2014. PECA: a novel statistical tool for deconvoluting time-dependent gene expression regulation. *J. Proteome Res.* 13(1):29–37
172. Cheng Z, Teo G, Krueger S, Rock TM, Koh HW, et al. 2016. Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol. Syst. Biol.* 12(1):855–855
173. Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, et al. 2016. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* 12(7):109
174. Yuan G-C, Cai L, Elowitz M, Enver T, Fan G, et al. 2017. Challenges and emerging directions in single-cell analysis. *Genome Biol.* 18(1):84
175. Budnik B, Levy E, Slavov N. 2017. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. bioRxiv 102681. **https://doi.org/10.1101/102681**
176. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Yosef N. 2017. The Human Cell Atlas. bioRxiv 121202. **http://dx.doi.org/10.1101/121202**
177. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44(D1):D457–62
178. Terfve CDA, Wilkes EH, Casado P, Cutillas PR, Saez-Rodriguez J. 2015. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun.* 6:8033
179. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, et al. 2006. COPASI–a COmplex PAthway SImulator. *Bioinformatics* 22(24):3067–74
180. Angermann BR, Klauschen F, Garcia AD, Prustel T, Zhang F, et al. 2012. Computational modeling of cellular signaling processes embedded into dynamic spatial contexts. *Nat. Methods* 9(3):283–89
181. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246(4926):64–71

182. Hillenkamp F, Karas M, Beavis RC, Chait BT. 1991. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* 63(24):A1193–1203
183. Eliuk S, Makarov A. 2015. Evolution of orbitrap mass spectrometry instrumentation. *Annu. Rev. Anal. Chem.* 8:61–80
184. Meier F, Beck S, Grassl N, Lubeck M, Park MA, et al. 2015. Parallel accumulation-serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* 14(12):5378–87
185. Graumann J, Hubner NC, Kim JB, Ko K, Moser M, et al. 2008. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol. Cell Proteom.* 7(4):672–83
186. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5(7):621–28

1.3 Interactomics

Since the invention of the tandem affinity purification (TAP) tag [Rigaut et al., 1999, Puig et al., 2001], which enabled the purification of generic proteins, MS has been successfully employed to identify interaction partners of many tagged proteins of interest. Samples are subjected to MS analysis after two successive purification steps and interactions are directly derived from the resulting list of identified proteins. Large-scale studies uncovered the interactome of the prokaryote *E. coli* [Butland et al., 2005] and the eukaryote yeast [Gavin et al., 2006], showing the competitiveness of the MS platform compared to the established genomic yeast two-hybrid (Y2H) system [Uetz et al., 2000] which had been used to generate PPI networks of even larger scale [Stelzl et al., 2005].

With the advent of quantitative MS, the paradigm shifted away from relying on purification and identification [Schulze and Mann, 2004] towards enriching the sample and subjecting the enriched sample to quantitative analysis [Keilhauer et al., 2015] (see Figure 1.4). Several large scale human PPI networks have been assembled using a quantitative analysis of MS data, including [Hein et al., 2015], where proteins were tagged with green fluorescent protein (GFP) using bacterial artificial chromosome (BAC) recombineering [Hubner et al., 2010], and the BioPlex project [Huttlin et al., 2015, 2017] which employed FLAG-HA tags.

The central challenge in the analysis of quantitative pull-down experiments is distinguishing false-positive background binders from real interactors [Nesvizhskii et al., 2007]. While initially simple fold-change cutoffs have been used, most studies now employ statistical testing and apply a significance cutoff to determine interactors. A popular choice is the t-test for comparing the means between a bait and a control pull-down. One down side of the t-test is its sensitivity to proteins with low-variance and small mean difference. Such proteins would be assigned a high p-value, despite the expectations set by the experimental enrichment that true interactors should exhibit a large effect size. In order to remove proteins with small effect size, either an additional fold-change cutoff can be applied [Kloet et al., 2016] or the t-test can be adjusted. The s_0-modified t-test [Tusher et al., 2001], initially developed for microarray analysis, smoothly controls the contributions of the effect size on the adjusted p-value [Keilhauer et al., 2015]. When testing multiple hypotheses, in this case one t-test for each protein in the dataset, a significance cutoff should not be applied directly to the p-values,

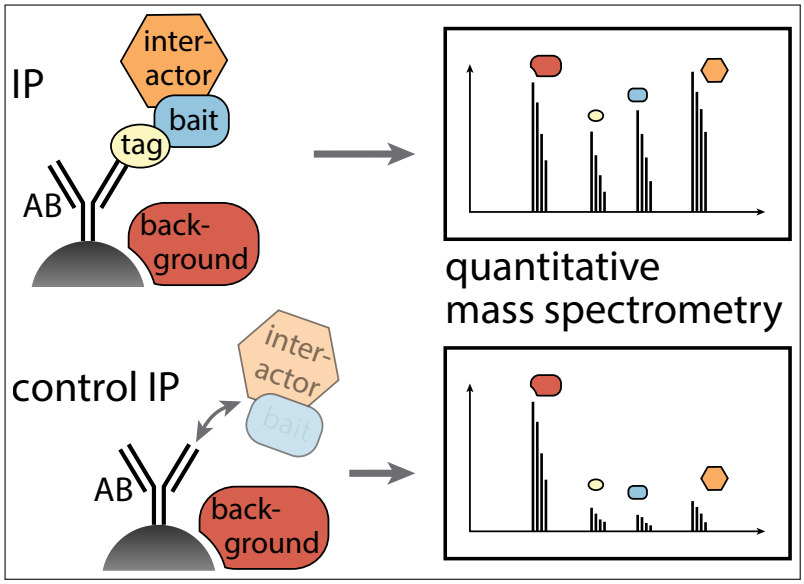


Figure 1.4: The affinity-enrichment-MS workflow utilizes quantitative proteomics to compare an enriched pull-down sample to a control sample. Adapted from [Hein et al., 2013].

which will lead to inflated overall false-positive rates. Instead, a scheme to control the false discovery rate (FDR) should be applied, such as the Benjamini-Hochberg correction [Benjamini and Hochberg, 1995]. A robust alternative, using permuted data to estimate the FDR, is implemented in the Perseus platform [Tyanova et al., 2016]. Alternatives to the *t*-test include fitting a mixture model to capture the distributions of true and false positives [Choi et al., 2011] or using a Naive Bayes classifier on semi-quantitative features derived from the data, including co-occurrence patterns of proteins in multiple bait pull downs [Huttlin et al., 2015].

1.3.1 Protein-protein interaction network databases

The networks generated by the community effort to uncover the interactomes of various organisms using genomic and proteomic technologies are stored in a zoo of PPI data bases. Any researcher interested in studying the interactome accesses available data through one of these databases. Each database has a different focus, such as a specific organism, the integration of multiple data sources or manual content curation. The STRING [Szklarczyk et al., 2015] database is one of the most popular choices because

of its easily accessible web interface and its inclusion of not only physical interactions but also non-physical interactions, such as co-expression, gene fusion or literature co-occurrence, and its wide coverage of organisms. Another PPI resource which additionally provides large-scale genetic interactions is the BioGRID [Chatr-Aryamontri et al., 2017]. Other databases, such as IntAct [Orchard et al., 2014] aggregate data from the community and focus on standardized data access. Organism specific data bases, including the Human Protein Reference Database (HPRD) [Keshava Prasad et al., 2009], only report manually curated information on human proteins, but their interactive data access, supplemented by additional tools for e.g. phosphorylation motif finding or, pathway annotations make them a useful resource for small scale studies.

The popularity of manually curated interaction resources is mainly driven by the large number of false-positive interactions plaguing large-scale databases. Via automatic curation of interactions, some databases report confidence scores for each interaction which distinguish well supported high confidence interaction from putative low-confidence interactions [Szklarczyk et al., 2015]. A number of PPI databases are specialized in aggregating and curating interaction networks and providing confidence scores [Yosef et al., 2009, Alanis-Lobato et al., 2017]

Popular network resources that are rarely used for large-scale analysis but rather for annotating and validating results include the KEGG [Kanehisa et al., 2016] and Reactome [Fabregat et al., 2016] pathway databases, and the CORUM [Ruepp et al., 2009] database of mammalian protein complexes.

1.3.2 Analysis of protein-protein interaction networks

One can distinguish two different classes of analyses on PPI networks. The first being concerned with understanding the topology of the interactome, the second utilizing PPI networks to interpret other large-scale -omics data.

In order to understand the topology of a network, it is useful to examine the basic properties of its nodes. Counting the number of neighbors of any node i gives rise to its degree k_i. In the same way a node degree distribution can be obtained for the network. It has already been observed, that complex networks, such as cellular PPI networks, often exhibit a specific power-law node degree distribution [Albert and Barabási, 2002]. If the node degree distribution matches the power-law

$$P(k) \sim k^{-\gamma} \tag{1.1}$$

Box 1.1: Graph notation

An undirected graph G is defined as a set of vertices $V = \{1, \cdots, n\}$ and edges $E = \{(i,j)$ if i,j are connected$\}$. If two vertices i and j are connected in the graph, an edge (i,j) is added to E. Due to the undirected nature of the graph the edges (i,j) and (j,i) are equivalent. Physical protein-protein interaction networks [Stelzl et al., 2005] can be modelled as an undirected graph where the vertices are the proteins and the edges represent their interaction.
In directed graphs the notation remains identical, but the existence of (i,j) does not imply the existence of the opposite edge (j,i). Such networks are well suited to model networks with directional flow of information, such as protein-DNA interaction networks [Johnson et al., 2007], where transcription factors initiate transcription on the DNA and not vice versa.
Simple graphs allow for only a single edge between each pair of nodes. The extension of a simple graph to allow for multiple edges per node is called a multigraph. In this thesis multigraphs will use the same notation as simple graphs except of edges being denoted as triplets (i,j,k) where i,j are the source and target nodes of the interaction and k provides an unique identifier distinguishing the edge from the other edges between the same nodes. Multigraphs can also be used to model a graph containing different types of interactions [Bensimon et al., 2012, Szklarczyk et al., 2015].

the network exhibits a so-called scale-free topology, which implies the existence a small number of hub nodes with high degrees connecting a large number of peripheral nodes with low degrees in a small-world network.

Extensions of the concept of node degrees to graphlet signatures, which more accurately capture local topology, have been applied to various challenges, such as clustering the network into modules [Milenković and Przulj, 2008], or aligning graphs between species with the aim of predicting novel interactions and transferring annotations [Malod-Dognin and Pržulj, 2015].

Differential expression (DE) analysis of omics data can be integrated with PPI networks [Cline et al., 2007]. Rather than performing independent DE analysis on each of the proteins, one can try to find differentially regulated local network modules by con-

sidering a number of proteins simultaneously [Ideker et al., 2002]. Furthermore, any set of proteins of interest can be extended with so-called subnetwork reconstruction methods [Yosef et al., 2011, Tuncbag et al., 2016]. Given a large-scale network a parsimonious subnetwork connecting all proteins of interest is derived, potentially including proteins which were not considered before. Several applications for the analysis of phosphoproteomic data are introduced in 1.4.1.

Given the size of large scale PPI networks (see Figure 1.5) the development of software tools is essential for handling and analyzing such networks. The landscape of software tools for the analysis of biological networks is diverse. Developer focussed libraries implemented in programming languages such as R [R Core Team, 2018] and Python [van Rossum, 1995] allow for the manipulation, analysis and visualization [Franz et al., 2015] of generic networks. For the analysis of biological networks, the open-source Cytoscape software [Shannon et al., 2003] has become one of the most popular tools. On the one hand, it provides users with a graphical user interface and rich visualization options and on the other hand, allows developers to extend Cytoscape through a plugin system [Maere et al., 2005]. Commercial tools, such as IPA [Quiagen Inc.], focus on providing higher quality annotations and expert knowledge stored in proprietary databases.

1.4 Phosphoproteomics

An estimated third of eukaryotic proteins being phosphorylated [Cohen, 2000] makes phosphorylation one of the key PTMs. In an active balance, kinases and phosphatases rapidly and reversibly modify proteins in order to alter their function [Hunter, 1995]. While in signaling, phosphorylation often acts in an activating or deactivating manner [Macek et al., 2009], it can modulate almost all protein functions, including localization, half-life and interactions [Cohen, 2000]. Disruptions of phosphorylation-mediated signaling leads to diseases, such as cancers and conversely makes kinases and phosphatases attractive drug targets [Ventura and Nebreda, 2006].

Since the detection of the first phosphorylation site on vitellin in 1906 by Levene and Alsberg, phosphorylation sites have been studied one site at a time. However, only an unbiased and comprehensive measurement of the phosphoproteome would allow the understanding of the effects of protein phosphorylation at a systems level. MS provides

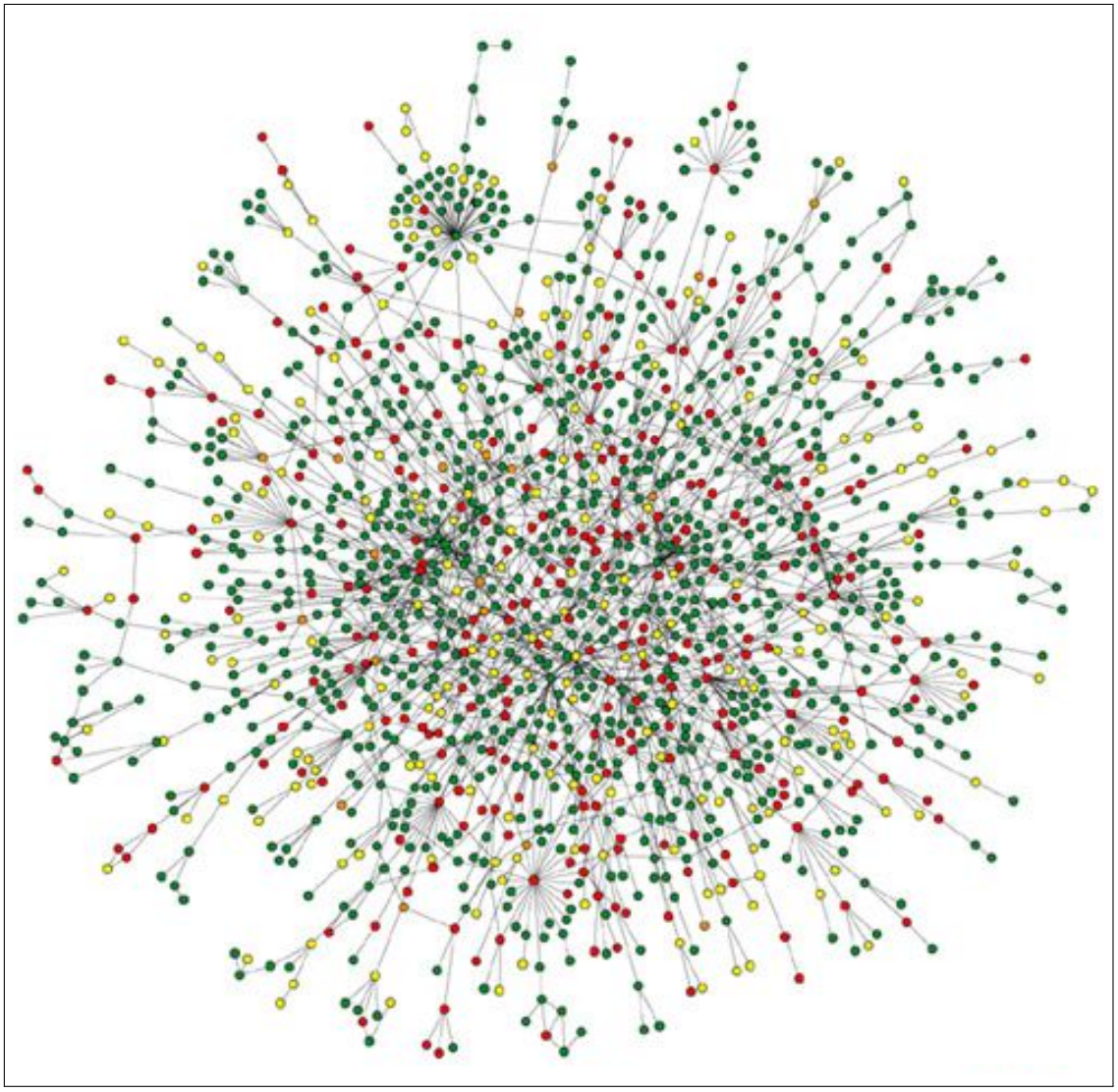


Figure 1.5: Node-link visualization of a large cluster of proteins from a yeast PPI network generated using the Y2H method [Uetz et al., 2000]. Protein nodes are colored according to the phenotype observed when removed (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown). Adapted from [Jeong et al., 2001].

an ideal platform to study PTMs such as phosphorylation which introduce a detectable mass shift at their modification site. The main challenges are not only to measure the phosphoproteome in a reproducible, high-throughput and deep-coverage manner, but also to adapt and extend the subsequent statistical analysis.

Due to the low stoichiometry of phosphorylation [Riley and Coon, 2016], phosphopeptides are hard to detect in regular proteomics experiments. Therefore, phosphoproteomic experiments are optimized to increase the coverage of phosphopeptides. First, digestion can be improved by using additional restriction enzymes which alleviate issues with suboptimal trypsin cleavage close to phosphorylated amino acids [Wiśniewski and Mann, 2012]. Second, a phosphopeptide enrichment step greatly increases the chances to detect low abundant phosphopeptides (see Figure 1.6). For smaller scale experiments, antibodies targeting phosphorylated residues can be enriched for by immunoprecipitation (IP) [Grønborg et al., 2002]. TiO_2 enrichment is better suited for large-scale experiments. Here, the oxygen in the phosphoryl group interacts with the metal oxid matrix [Thingholm et al., 2006]. Alternatively, immobilized metal affinity chromatography (IMAC) [Villén and Gygi, 2008] exploits the affinity of the negatively charged phosphate group on the peptide, to the positively charged metal cations in the column. By utilizing extensive prefractionation the coverage of phosphopeptides can be increased further. The downside to prefractionation is the increased amount of labor and measurement time required for the analysis. Recent streamlined protocols in 96-well format coupled with single-shot LC-MS enable high-throughput measurements of a large number of phosphoproteomic samples [Humphrey et al., 2015, 2018]. For the quantification of phosphoproteomics experiments, label-free as well as metabolic and isobaric labeling have been employed successfully [Hogrebe et al., 2018].

Compared to the identification of unmodified peptides from MS^2 spectra, the identification of the modified phosphopeptide and the localization of the modification on the peptide sequence require special considerations [Potel et al., 2018, Sinitcyn et al., 2018a]. In a database search framework, extending the sequence database with modified sequences allows for the identification of modified sequences. The mass shift on parts of the fragment ion series alongside modification-specific neutral losses and diagnostic peaks are considered when generating theoretical spectra. A final localization probability can be derived from the scores assigned to the theoretical fragment spectra with the different possible localizations [Beausoleil et al., 2006, Cox et al., 2011].

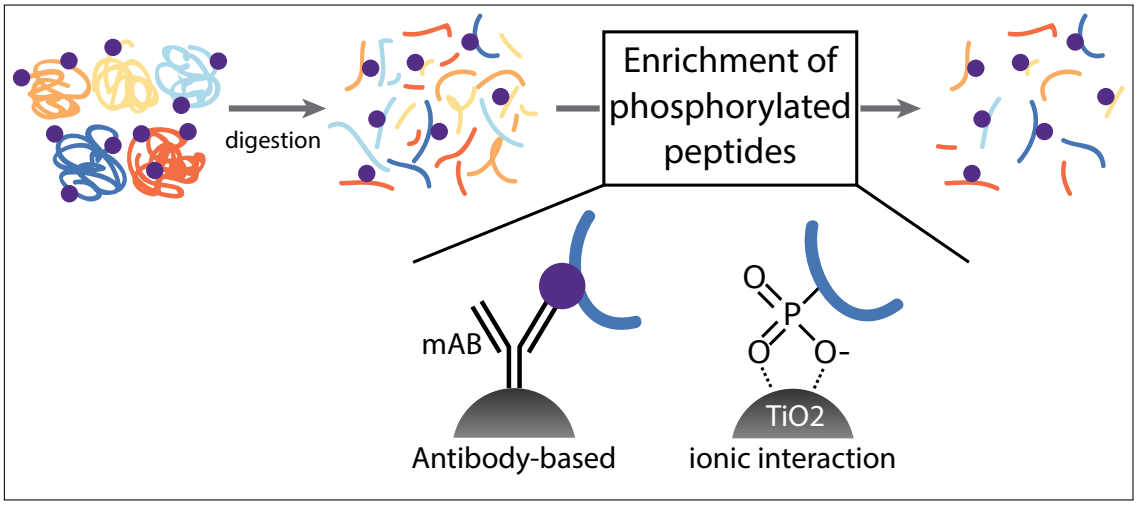


Figure 1.6: The phosphoproteomics workflow includes an additional phosphopeptide enrichment step prior to MS analysis. Adapted from [Hein et al., 2013].

When complementary proteome data is available stoichiometry can be derived from the measured intensities by calculating the ratios between the modified and unmodified peptide [Cox and Mann, 2008, Olsen et al., 2010, Sharma et al., 2014].

The identification of large numbers of phosphorylation sites has been successfully used for evolutionary studies. Phosphorylation sites were found to be conserved when occurring in structured regions of the protein, and rapidly evolving when located in unstructured regions [Collins, 2009]. To study the regulatory effect of phosphorylation, a quantitative analysis of the up to 50,000 detected sites detected by MS is required [Sharma et al., 2014]. The first of the two core challenges in the analysis phosphoproteomic data is the increased variability inherent to peptide-level data. Protein-level proteome analyses benefit from peptide to protein aggregation, leading to higher confidence data [Sinitcyn et al., 2018a]. The second challenge is the translation of the observed changes in phosphorylation patterns into mechanistic insights, since in general, phosphorylation sites can exhibit excitatory, inhibitor or even combinatoric effects on the function of the substrate [Macek et al., 2009].

While many phosphoproteomic studies still pursue analysis approaches taken from proteome analysis, such as identifying differential expression or clustering on a phosphorylation site level [Olsen et al., 2010, Sharma et al., 2014], a number of tools have been developed to specifically address the challenges of phosphoproteomic analysis.

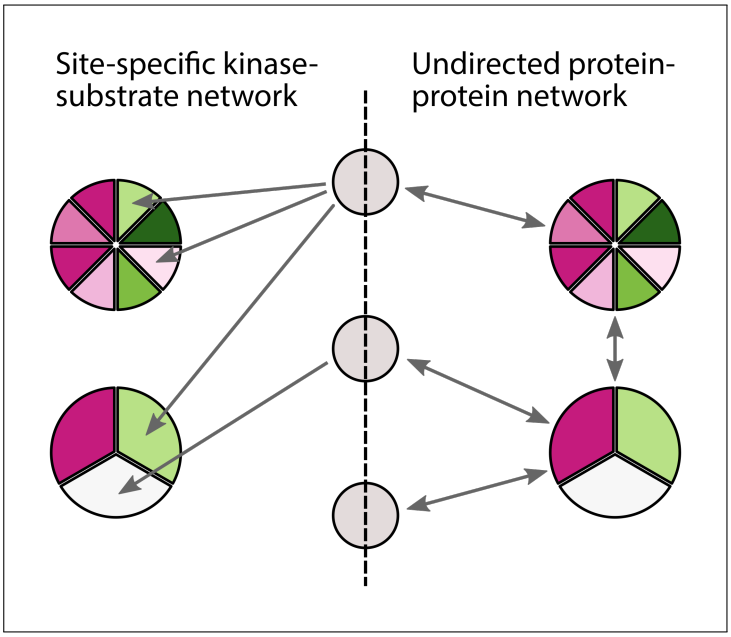


Figure 1.7: Kinase-substrate networks can be represented as either directed, site-specific networks which naturally arise from kinase-substrate interactions or kinase motifs, or as undirected protein-level interactions modeled after large-scale PPI networks. Adapted from [Rudolph and Cox, 2018].

One core concept that is exploited by most of these analyses it the relationship between kinases and their substrate, which allows to restate the aims of the analysis. Instead of focusing or the individual phosphorylation sites, the aim is to identify regulated kinases from large-scale phosphoproteomic data. Changes in the phosphorylation levels of the substrates imply corresponding changes in the activity of the kinase.

1.4.1 Kinase-substrate networks and kinase activities

Prerequisite for a kinase-centric analysis of phosphoproteomic data is the assignment of kinases to their specific substrates. To do so, phosphorylation events can be naturally described as a directed interaction between a kinase and a specific site on the substrate (Figure 1.7) forming a site-specific kinase-substrate network. Only a limited number of such site-specific interactions are available in databases such as PhosphoSitePlus [Hornbeck et al., 2015]. However, in order to obtain a comprehensive picture on kinase activity, a high coverage of kinases and sites would be required.

Due to the difficulties in uncovering novel interactions experimentally, a number of computational tools address this issue. The statistical analysis of the peptide sequences identified in large-scale phosphoproteomics experiments allows for the identification of phosphorylation motifs which can serve as a proxy for a kinase [Schwartz and Gygi, 2005]. The NetPhorest tool [Miller et al., 2008] integrates motif finding with the phylogeny of kinases to pinpoint the kinase family of a phosphorylation site. Combined with the network context modeling approach of NetworKIN [Linding et al., 2008], the KinomeXplorer [Horn et al., 2014] platform, integrates both tools to predict human kinase-substrate interactions, which can be used to supplement kinase-substrate networks from literature.

Most tools for the scoring of kinase activities from phosphoproteomic data have a common approach. Using the kinase-substrate assignments from the network, an activity score is calculated for each kinase/kinase family/motif from the observed phosphorylation changes on the assigned sites. The KSEA [Casado et al., 2013] method includes a variety of scores, including a Z-score and its associated p-value, which performs well for relative comparisons between conditions [Hernandez-Armenta et al., 2017]. For comparisons within the same conditions KARP [Wilkes et al., 2017] ranks the kinases based on their contribution to the total observed phosphorylation. Computationally more involved approaches try to optimize binding affinity between kinases and substrates (IKAP [Mischnik et al., 2015]), or even derive an entire logic model from a phosphoproteomic screen of kinase inhibitors [Terfve et al., 2015].

The requirement for site-specific interactions and the focus on deriving the enzymatic activity of kinases from the data limits the scope such methods. While the phosphorylation itself is mediated by the kinases, non-kinase proteins contribute to signaling by enabling signal transduction through complex forming and scaffolding. The PHOTON [Rudolph et al., 2016, Rudolph and Cox, 2018] method devises a more general signaling functionality score which allows it to use large-scale PPI networks to perform the analysis. The undirected, site-agnostic nature of such networks (Figure 1.7) reduces the specificity of the approach, but dramatically increases the quality of the network and the amount of data that can be utilized in the analysis. The PHOTON score itself is calculated from the average phosphorylation changes observed on the neighbors of any protein in the network. A permutation-based FDR approach determines which proteins have significant signaling functionality scores.

1.5 Co-expression analysis

Co-expression analysis is seeking to identify functional modules of genes or proteins from omics experiments with large numbers of samples. First, a co-expression network is constructed directly from the data, in which the node are connected by edges with weights according to their co-expression. The extent of co-expression between genes is most often measured using correlation [Langfelder and Horvath, 2008], or mutual information [Margolin et al., 2006]. Finally, the clustering of the network reveals the co-expression modules.

Choosing the appropriate co-expression metric is central to the analysis. Correlation measures the linear dependence between a pair of genes, while mutual information can capture non-linear relationships. On the other hand, the estimation of mutual information from non-discrete data, such as protein expression measurements is more challenging [Kraskov et al., 2004]. Biweight midcorrelation provides a robust alternative to Pearson's correlation. For vectors $x = (x_a), a = 1, 2, \ldots, m$ and y it is defined as

$$\text{bicor}(x, y) = \sum_{a=1}^{m} \tilde{x}_a \tilde{y}_a \tag{1.2}$$

$$\tilde{x}_a = \frac{(x_a - \text{med}(x)) w_a^{(x)}}{\sqrt{\sum_{b=1}^{m} \left[(x_b - \text{med}(x)) w_b^{(x)} \right]^2}} \tag{1.3}$$

$$w_a = (1 - u_a^2)^2 I(1 - |u_a|) \tag{1.4}$$

$$u_a = \frac{x_a - \text{med}(x)}{9\,\text{mad}(x)} \tag{1.5}$$

Biweight midcorrelation was found to outperform Pearson-, Spearman correlation, and mutual information measures for co-expression network construction in synthetic and real data sets [Song et al., 2012].

A transformation of correlation coefficients ρ_{ij} into the range $[0, 1]$ is required in order to obtain the adjacencies a_{ij} that form the co-expression network. A *signed* transformation retains the direction of the correlation, while the *unsigned* transformation does not. Additionally, raising the adjacency measure to the power β acts as a soft-threshold

on the network.

$$a_{ij}^{\text{signed}} = \left(\frac{1+\rho_{ij}}{2}\right)^{\beta} \tag{1.6}$$

$$a_{ij}^{\text{unsigned}} = |\rho_{ij}|^{\beta} \tag{1.7}$$

The choice of β can be informed by looking at network properties, such as average connectivity, or scale-freeness (see Equation 1.1) of the network. A scale-free fit index can be derived from regressing on the log-transformed node-degree distribution [Zhang and Horvath, 2005]. For higher powers of β, average connectivity will reduced while the scale-free fit index should improve.

Once the co-expression network is constructed a cluster dendrogram can be calculated using hierarchical clustering. Instead of the often used euclidean distance, the topological overlap distance $1 - t_{ij}$ is applied. The topological overlap measure (TOM) t_{ij} measures the distance between nodes in a network based on their first-degree neighbors [Ravasz et al., 2002].

$$t_{ij} = \begin{cases} \frac{l_{ij}+a_{ij}}{\min\{k_i,k_j\}+1-a_{ij}} & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{1.8}$$

$$l_{ij} = \sum_u a_{iu} a_{uj} \tag{1.9}$$

$$k_i = \sum_u a_{iu} \tag{1.10}$$

Modules can be extracted from the cluster dendrogram using a simple distance cutoff, or using automated cluster detection by dynamic branch cutting [Langfelder and Horvath, 2008]. Based on the guilt by association (GBA) principle [Oliver, 2000] module members can be characterized and novel associations between members [Lee et al., 2004] can be established. The biological relevance of the entire module can be assessed using annotation enrichment analysis using Fisher's exact test.

In order to make comparing modules and relating them to external traits, such as phenotype or clinical data easier, each module can be represented by an module eigengene [Zhang and Horvath, 2005]. The eigengene is derived from a principal component analysis (PCA) of the quantitative profiles of all module members, and has a quantitative profile that summarizes the module behavior. Subsequently, module to phenotype links can be established by calculating the correlation between the eigengene and the

observed phenotype. A heatmap of the correlations between the module eigengenes and the phenotypes can provide a highly reduced view of large datasets.

Chapter 2

Manuscripts

2.1 A network module for Perseus

We recently developed and published the popular Perseus software for the downstream statistical analysis of proteomics data [Tyanova et al., 2016]. The following manuscript introduces PerseusNet, the network module of Perseus. PerseusNet was devised to fulfill the computational needs of proteomics researchers wishing to accomplish network analysis of their data. While it is extensible through a new plugin application programming interface, and hence any network analysis functionality can be implemented, most tools needed for proteomics research and connecting it to generic network analysis platforms are included in the software. Dedicated activities for analyzing AP-MS datasets and phospho-proteomics experiments in the context of kinase-substrate networks belong to the basic infrastructure of PerseusNet. PerseusNet is extensible through a plug-in architecture in a multi-lingual way, integrating scripts in C#, Python and R, which allows for the incorporation of a plethora of existing scripts and programs from the network community.

My contribution was the implementation of PerseusNet, including the required internal data structures and algorithms, as well as the activities that allow for loading, processing, analysing and visualizing proteomics, phosphoproteomics and interactomics data. I additionally designed and implemented a number of software modules (PluginInterop, perseuspy, PerseusR) which enable the interoperability between Perseus and generic data science tools, as demonstrated by the integration of the WGCNA [Langfelder and Horvath, 2008] and PHOTON [Rudolph et al., 2016] into

Perseus.

Jan Daniel Rudolph and Jürgen Cox. A network module for the Perseus software for computational proteomics facilitates proteome interaction graph analysis. *bioRxiv*, page 447268, October 2018. doi: 10.1101/447268. URL https://www.biorxiv.org/content/early/2018/10/18/447268?rss=1

A network module for the Perseus software for computational proteomics facilitates proteome interaction graph analysis

Jan Rudolph[1] and Jürgen Cox[1,*]
[1]Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.
*Correspondence: cox@biochem.mpg.de

ABSTRACT

Proteomics data analysis strongly benefits from not studying single proteins in isolation but taking their multivariate interdependence into account. We introduce PerseusNet, the new Perseus network module for the biological analysis of proteomics data. Proteomics is commonly used to generate networks, e.g. with affinity purification experiments, but networks are also used to explore proteomics data. PerseusNet supports the biomedical researcher for both modes of data analysis with a multitude of activities. For affinity purification, a volcano plot-based statistical analysis method for network generation is featured which is scalable to large numbers of baits. For posttranslational modifications of proteins, such as phosphorylation, a collection of dedicated network analysis tools helps elucidating cellular signaling events. Co-expression network analysis of proteomics data adopts established tools from transcriptome co-expression analysis. PerseusNet is extensible through a plug-in architecture in a multi-lingual way, integrating analyses in C#, Python and R and is freely available at http://www.perseus-framework.org.

INTRODUCTION

The study of complex systems[1] is concerned with the question of how the relationship between the parts of a system give rise to its collective behavior. Complex systems often generate emergent properties[2] which are not contained in an obvious way in its parts. Examples of such networks range over all disciplines of science, including the study of social media networks[3], scientific collaboration networks[4] and the human brain and its interconnected neurons as a particularly interesting one. The interactions between the components of a complex system define a network of connections consisting of nodes and edges. Much of the relevant content is concealed in the network constructed from these interactions and is not visible in the components themselves. For instance, the brain connectome[5] is believed to make us who we are and not the cellular content of the brain[6]. Similarly, the observation of cellular concentrations of biomolecules without considering their interaction would provide a limited picture that ignores potential emergent properties of the biomolecular complex system. Hence it is mandatory to study biological systems, such as cellular concentrations of biomolecules, in the framework of network biology[7].

At a fundamental level, all network connections between the cellular biomolecules are biochemical reactions and their specification in biochemical pathways together with their subcellular spatial distribution would provide complete knowledge about the biological network state of the cell. This collective network of all biochemical reactions contains all metabolic reactions, the signaling cascades, gene regulatory networks and all complex-forming non-covalent interactions between molecules, as for instance protein-protein interactions. Due to the limitations of experimental and computational methods to map out this interaction network, we often obtain only partial knowledge about the complete biochemical reaction network from experiments. Networks are however not limited to describing fundamental physico-chemical interactions between biomolecules. For instance, in a gene co-expression network analysis[8] one looks for similarity of expression patterns of gene products over many samples. Strongly correlated expression implies that these genes have some kind of non-physical interaction, e.g. they are part of the same transcriptional regulatory program or they share membership in the same pathway or protein complex. However, the exact relationship in terms of biochemical reactions remains unknown with these and other techniques. Hence, in these cases, networks describe a coarser grained level of detail, in which relationships between molecules are not necessarily biochemical reactions, but of a more general kind.

Computational proteomics is a mature data science that copes well with the large amounts of data produced in mass spectrometry experiments[9]. Perseus is an established framework for the downstream bioinformatics analysis of quantitative proteomics data[10,11]. The initial version of Perseus provided a

comprehensive framework and set of activities to analyze data matrices originating from quantitative proteomics in a workflow environment. The main idea behind Perseus is to enable the researchers in biomedical sciences to perform the data analysis themselves. Here we describe how we extend this program to the analysis of biological networks in the context of proteomics. While cytoscape[12] exists as the de-facto standard for network analysis and visualization, many proteomics-specific tasks for the generation and analysis of networks are lacking from this framework, as well as workflow navigation. PerseusNet fills this gap and enables non-computational experts to perform complete network-based analysis of their data. We explicitly do not want to re-invent existing methods and algorithms. Instead we designed an extensible framework that integrates with existing tools, like cytoscape, and interoperates with existing code and scripts from the network analysis community that were written in diverse languages, like Python and R. The data structures within Perseus that hold the networks were set up in a way that facilitates studying dynamic changes in networks and finding differential network properties over complex experimental designs. Side-by-side analysis of networks with data matrices in a common workflow environment allows for a seamless transition between matrix-centric and network-centric approaches.

In the following we start with a general description of the new network framework in Perseus, including how it enables multilingual programming and usage of code resources from R and Python. Then we introduce the new volcano-plot based analysis workflow scalable to large affinity purification-mass spectrometry (AP-MS) datasets. We describe how general and more specifically, large-scale protein-protein interaction (PPI) networks are handled and curated in Perseus. A section on the analysis of posttranslational modification (PTM) induced networks, like kinase-substrate relationships for phospho-proteomics is next. Finally, we cover co-expression analysis in Perseus and its applications to clinical proteomics.

RESULTS

WORKFLOW-BASED BIOLOGICAL NETWORK ANALYSIS

PerseusNet was devised to fulfill the computational needs of proteomics researchers wishing to accomplish network analysis of their data. While it is extensible through a new plugin application programming interface (API), and hence any network analysis functionality can be implemented, most tools needed for proteomics research and connecting it to generic network analysis platforms are included in the software (Fig. 1). Dedicated activities for analyzing AP-MS datasets and phospho-proteomics experiments in the context of kinase-substrate networks belong to the basic infrastructure of PerseusNet. The most common standard data formats are supported as input. An extended multi-language plugin API allows leveraging many existing tools in the analysis workflow. As an important example, co-expression clustering tools are integrated in this way.

To accommodate PerseusNet, we extended the Perseus framework with a new data type termed network collection (Fig. 2) that represents a set of one or more networks which are analyzed jointly in the workflow. Different networks within the same network collection can, for instance, represent networks derived from different individuals (patients), experimental conditions or biological replicates. All information in the network collection is organized in data tables, leveraging the existing augmented data matrix[10] in Perseus. General information on the networks in the collection are stored in the networks table, where each row represents an individual network. Here, sample-related annotations, such as calculated global network properties, can be stored to enable their usage in analysis activities operating on a network collection. For instance, if the samples correspond to different patients, the networks table can hold patient-specific information as derived from patient records or questionnaires. These variables can then be used as independent or confounding factors in statistical analysis of the networks.

The nodes and edges of each individual network are stored in a pair of separate tables. The nodes table further describes the entities in the network, while the edges table provides details on the connections between the entities. The entities in the nodes table can be annotated with local network properties, such as the node degree. In case the entities correspond to proteins, biologically meaningful annotations could include membership in gene ontology terms, pathways or protein complexes. Similarly, edges can be annotated in the edges table with properties of pairwise relationships between proteins, as for instance interaction confidence measures. All of these properties are then accessible to the network analysis tools. Furthermore, all mentioned tables can be sorted and searched, allowing all information to be browsed and inspected intuitively. Internally, a graph data structure for each network enables the efficient execution of

graph algorithms. We did not aim to include generic graphical representation of networks as node-link diagrams, since this can be achieved in other tools such as Cytoscape, for which we provide simple adaptors for the transfer of networks. However, several activities include specialized visualizations tailored to specific analyses.

In Perseus, all data analysis steps are performed within a graphical workflow. (See Supplementary Fig. 1.) Enabled by the newly implemented network collection, the Perseus workflow is now capable of all import, processing, and analysis steps in the side-by-side analysis of expression matrices and networks. All data that is imported into Perseus is represented as a separate entity in the workflow. Any matrix or network undergoing a processing step is not modified in-place but rather becomes a new entity that gets connected to the original data in the workflow. By inspecting both, input and output data, every step in the analysis is traceable and easily understood. Certain processing steps allow for the transformation of matrices into networks and vice-versa, or the mapping of data between the two. As a result, any analysis performed in Perseus, potentially including several side-by-side processing steps of networks and matrices, always remains transparent to the user.

MULTILINGUAL PLUG-IN ACTIVITIES

The network collection data structure (Fig. 2) and the extended Perseus workflow provide the foundation for enabling various network analyses, many of which are available in Perseus. In general, networks either originate from external sources, or are created in a data-driven manner from within the workflow. To facilitate the import of external networks into the workflow, we implemented parsers for standard network formats, such as edge table (.tab|.txt|.csv), GraphML[1] (.gml), Cytoscape's simple interaction format[2] (.sif) and D3js's JSONgraph[3] (.json) which enable loading interactions from most popular network databases, including STRING[13], BioGRID[14], IntAct[15], CORUM[16] and PhosphoSitePlus[17]. Furthermore, specific quantitative expression data, such as AP-MS drives the creation of novel protein-protein interaction networks, and phospho-proteomics datasets allow for a more detailed view or construction of kinase-substrate relationship networks. Specialized visualizations of such networks are provided (see later sections), which allow for an intuitive visual inspection of the results of the analysis. Perseus is not limited to physical interaction networks: co-expression clustering provides a powerful alternative to regular hierarchical clustering for expression proteomics studies. Finally, any network collection can be exported from the workflow in a plain text file format for sharing or use in any other external tools, such as Cytoscape. In order to accommodate these new capabilities in the Perseus plugin system, we extended the Perseus plugin API with new programming interfaces for the network collection and other associated data types, as well as the respective import, processing, and analysis interfaces. (See Supplementary Fig. 2.) This fully-featured API is available to all developers wishing to extend Perseus' functionality with plugins. All analyses presented in this manuscript adhere to the new API.

In-order to better leverage the existing network analysis ecosystem, we additionally implemented a new mode of interoperability between Perseus and external tools (Fig. 3). The PluginInterop project enables this functionality, and allows the user to run external tools from within the Perseus workflow, most prominently scripts written in the popular R and Python languages. Open-source companion libraries for R[4] and Python[5] provide utilities for interfacing with Perseus. As a result, network analysis tools originally implemented in external tools can run from within the Perseus workflow with only minor adjustments. The implementations of the PHOTON and WGCNA plugins presented in the manuscript are based upon PluginInterop and its companion libraries. Instructions for interested developers on how to write scripts for Perseus or how to adapt existing tools can be found on the PluginInterop website[6]. In the following sections, we will present a number of network analyses which are now implemented in Perseus, with focus on their application to different types of proteomics data.

AFFINITY ENRICHMENT MS INTERACTOMICS

Affinity purification or enrichment coupled to MS analysis has become a powerful tool for interrogating PPIs[18,19]. It is able to provide not only a detailed view on proteins of interest, but it can also determine the basic building blocks for the assembly of large-scale protein-protein interaction networks[20,21]. Historically, protein complex members were detected by subjecting the sample to a series of purification steps followed by MS

[1] http://graphml.graphdrawing.org/
[2] http://manual.cytoscape.org/en/stable/Supported_Network_File_Formats.html
[3] http://jsongraphformat.info/
[4] PerseusR, https://github.com/jdrudolph/PerseusR
[5] perseuspy, https://github.com/jdrudolph/perseuspy
[6] https://github.com/jdrudolph/PluginInterop

identification. With the advent of quantitative MS, detecting even transient interactions has become possible by relying not only on the identification itself, but instead on quantitative information. The sample is not purified, but only enriched for the protein of interest and its interaction partners and then subjected to MS quantification[22].

Confidently identifying bona-fide interactions and distinguishing them from background binders, arising from off-target binding or contamination, requires data analysis of replicate case and control measurements. Compared to purely fold change-based methods, statistical tests provide a powerful way to compare case and control samples by calculating a test statistic and an associated p-value and limit the number of false-positives. For visual inspection of the results, the (negative logarithm of the) p-value can be plotted against the size of the effect, i.e. the difference between the means of logarithmic abundances, in a so-called volcano plot. Since one statistical test is performed for each protein, which amounts to a large number of tests performed simultaneously, the significance level needs to be adjusted to avoid increased numbers of false positives due to the multiple hypothesis testing problem[23]. A popular strategy to adjust for multiple testing is to control the false discovery rate (FDR), which can be achieved by permutation-based methods. Furthermore, in the volcano plot method it is necessary to define the functional form of the curves that separate significant from non-significant hits, either by straight lines, or in a more sophisticated way, introduced in the significance analysis of microarrays (SAM) method[24], by modifying the t-test statistic with the background variance parameter s_0. This standard workflow is available in Perseus but becomes increasingly cumbersome for interaction screens with more than a handful of baits. Parameter values for s_0 and the FDR thresholds are often applied separately for each pulldown, inviting overfitting and cherry-picking, and also requiring results be subsequently combined manually.

We implemented the interactive multi-volcano plot (Fig. 4a) to analyze interaction screens with arbitrarily many baits and conditions simultaneously. Given the experimental design of the dataset, defined by baits and conditions, the analysis is applied to each experiment. For sufficiently large datasets, instead of dedicated control samples, an internal control can be assembled from the dataset for each pulldown consisting of pulldowns of other, unrelated baits. The results can be inspected through an interactive user interface. All volcano plots are displayed in the overview panel. A multi-functional detail panel shows more information on selected plots and provides zoom, protein selection and labeling options. If a single plot is selected, the volcano plot is shown in the detail panel. When two plots are selected, the t-test differences between the selected experiments are plotted against each other, highlighting changes in the enrichment of proteins between experiments (Fig. 4b). Additionally, all data can be browsed in tabular form, making it easily searchable and allowing for rich styling options. Known interactors or gene ontology annotations matching the experiment can be used to highlight proteins in the plot and can serve as a positive control for the adjustment of test parameters. All test parameters are controlled on a global level, effectively preventing overfitting and cherry-picking parameter values. We integrated the multi-volcano analysis into the new network module. Results from protein-protein interaction screens can be exported as network objects into the Perseus workflow. A specialized node-link visualization based on the open-source cytoscape.js library[25,26] with multiple layers of information, allows for easy interpretation of the results (Fig. 4c). A protein-protein interaction network that was newly created in this way can be integrated with existing networks, or exported in various formats using the functions available through the network module.

As an example application, we obtained pull-down experiments from reference[27], covering three baits in two different cell types. The filtered data set contained 2995 proteins. Using the new multi-volcano analysis (Fig. 4a), we obtained a PPI network with 134 nodes and 140 edges. The results were comparable to the original publication with overlaps between 55% (Ring1b ESC) and 91% (Bap1 ESC) for Class A interactions. Differences can be explained by the slightly different methodology used in this manuscript. We used the s_0-modified t-test with s0 set to 1.0, and FDRs of 0.01% and 0.2% for Class A and B, respectively, while the authors of reference[27] used individually chosen fold-change and p-value cutoffs for each experiment. Using the built-in visualization features, such as the enrichment between experiments, we identified several interactions that were conditional on the cell type (Fig. 4b). By annotating the newly created protein-interaction network with known complex interaction from CORUM and inspecting the resulting node-link network visualization (Fig. 4c), previously known and possibly novel interactions could be distinguished.

Further confidence in the existence of an interaction between a protein identified in a pulldown and the bait can be obtained by correlation analysis. The correlation

of the intensity profiles over many pulldowns with the bait intensity profile is reported in the output tables together with the volcano plot-derived significance of the interaction. When assembling the interaction network a threshold is applied to this correlation in order to define an additional class of interactions (Class C), which might not have been found by volcano plot analysis (Classes A and B). This workflow is especially appealing for interaction screens with a large number of bait proteins.

IMPORTING, CURATING AND PROBING LARGE-SCALE PPI NETWORKS

While protein interaction screens can uncover novel or condition-specific interactions, a wealth of detected and predicted interactions are already stored in protein-protein interaction databases[28]. Analyzing large-scale PPI networks jointly with other omics data has great promise. However, a major obstacle to performing systems-level analysis on these large-scale networks are lacking easy-to-use software solutions to transparently handle the processing and analysis of these networks. Many studies under-utilize the existing resources and mostly report the interactions of a single protein as an after-thought. In the following, we introduce the new network capabilities of Perseus to assemble, filter, and understand large-scale PPI networks, which lay the foundation for any network analysis.

The first task is assembling a high confidence interaction network. Many databases, such as STRING[13], BioGRID[14] or HIPPIE[29], allow researchers to download all interactions in a tabular format, which can be easily loaded into Perseus, even with sizes of up to few millions of interactions. Supplementary information on the interactions such as, but not limited to, the interaction type or a measure of confidence, remain available at each step in the subsequent data analysis. With this information, generalized interaction networks, such as STRING can be filtered by interaction type to generate a physical interaction network. Confidence measures often integrate diverse knowledge into a single score, derived from how often, and by which experimental technique, an interaction was detected, combined with more abstract measures, such as co-expression and literature co-occurrence of the interaction partners[13]. There are two approaches for interaction confidence aware network analysis (Fig. 5a). Applying a cutoff to the confidence score removes low-confidence interactions from the network, which is especially useful when applying methods that treat all interactions equally. The cutoff can be chosen according to the confidence score distribution and the targeted network size (Fig. 5b). Other methods operate on weighted networks and distinguish between interactions with high or low confidence. In this case the confidence scores can be used as an edge weight. In addition to static confidence scores, one can devise dynamic confidence scores from experimental data which reflect e.g. changes in abundance or localization of any of the interactors.

A deeper understanding of the network requires a different perspective in addition to the interaction-centric view. Any list of interactions can be converted into a network collection with a single click. A dedicated set of network-specific processing activities are now available. While processing the list of interactions the focus remains on the edges of the network. In the network view, the focus is shifted to the nodes. With the powerful identifier and data mapping mechanisms in Perseus, nodes are easily annotated with various annotations, such as gene ontology[30] (GO), or quantitative proteomics data. Any annotation can be subsequently used to filter the nodes of the network. One could, for example, extract a subnetwork of proteins associated with a specific GO category and their interactions from the large-scale network. Using the data mapping from e.g. deep proteomes of specific cell-lines or tissues, condition-specific subnetworks can be created.

Further understanding is gained by studying the intrinsic properties of networks. By calculating node degrees, corresponding to the number of neighbors of each node in the network, hub nodes can be distinguished from peripheral nodes. By analyzing the distribution of the node degrees in the network, global network properties, such as approximate scale-freeness[31] of the topology can be identified (Fig. 5c). Furthermore, intrinsic local network properties, like the node degree can be correlated with biological properties derived from protein annotations or experimental data. The proper construction of large-scale interaction networks and understanding their basic properties are central to the successful application of more specialized analyses such as the integration of such networks with PTM data.

NETWORK ANALYSIS OF PTM DATA

The MS-based study of PTMs is nowadays possible on a global scale for several types of modifications. The best known example is MS-based phosphoproteomics[32], which is a powerful tool for interrogating signaling events on a large scale. However, drawing conclusions directly from phosphorylation changes is challenging,

due to the mostly missing functional information on the inhibitory or excitatory action of a specific protein phosphorylation at a specific site. Network-based approaches for the analysis of phosphorylation data derive functional information on protein-level by interrogating the phosphorylation changes observed in the network neighborhood[33–35].

We implemented the popular kinase-substrate enrichment analysis[34] (KSEA) tool for predicting kinase activities in Perseus. Site-specific kinase-substrate networks (Fig. 6a) assign kinases to the experimentally observed phosphorylation sites. The core of the analysis is the calculation of a series of scores (mean, enrichment, Z-score, p-value, q-value) for each kinase, based on the quantitative phosphorylation changes of its substrates. These predicted kinase activities can be analyzed further to find e.g. differentially activated kinases or pathways. KSEA most often utilizes the curated kinase-substrate network from the PhosphoSitePlus database[17,36,37]. In order to extend the coverage of the network and thereby allow for the utilization of a larger fraction of the experimental data, the network can be supplemented with predicted kinase-substrate interactions from tools such as NetworKIN[38,39], or with low-specificity interactions derived from kinase target sequence motifs.

PHOTON[33], now implemented in Perseus, is an alternative approach to KSEA that calculates more broadly defined signaling functionality scores for any protein, rather than activities for kinases only. A data-annotated large-scale PPI network now serves as the input (Fig. 6a). The resulting signaling functionality scores for each experimental condition are based on the observed phosphorylation in the neighborhood of each proteins and are assigned a significance by a permutation-based FDR scheme. The scores can either be analyzed directly, to find proteins and pathways with e.g. differentially changing signaling functionality or utilized in a second step of the PHOTON pipeline, in which signaling pathways are automatically reconstructed from the large-scale network, that connect the proteins with significant signaling functionality.

The Perseus network module allows for performing both KSEA, and PHOTON analysis on the same experimental data[33] and a choice of networks[17,29]. Due to the differences in the utilized methodologies and the chosen networks, resulting scores will differ, but are easily compared with the analyses and visualizations provided by Perseus (Fig. 6b). Both tools support the analysis of datasets with multiple conditions, effectively transforming the peptide-level phosphorylation data into protein-level scores. The entire well established toolset for the analysis of protein quantification data can be applied to these scores, including hierarchical clustering, enrichment analysis[33] and time-series analysis[40].

We implemented an interactive visualization of kinase-substrate networks directly in Perseus (Fig. 6c) using the cytoscape.js library[25]. The visualization allows for the joint visual inspection of the networks, e.g. subnetworks reconstructed by PHOTON, and the quantitative phosphorylation data. Browsing the quantitative phosphorylation in a reduced and highly structured network view while also considering the signaling functionality scores, allows for the generation of hypotheses that explain the signal transduction mechanistically.

CO-EXPRESSION CLUSTERING AND CLINICAL DATA

When performing co-expression analysis, the correlation matrix between the proteins in the dataset describes a fully connected, weighted network, in which the weight on each edge denotes the correlation between the two proteins (Fig. 7a). Hence, the actual network usually remains implicit. A hierarchical clustering of the co-expression network can utilize the network neighborhood of each protein and integrate it into the similarity calculation[41]. The cluster dendrogram and the detected co-expression modules are then transferred back to the original data where their interpretation is equivalent to ordinary hierarchical clustering. In addition to the clustering, a representative expression profile for each of the clusters is generated, which is termed eigengene. This highly reduced view on the data can be correlated with clinical or phenotype data and clustered to gain a better understanding of the behavior of the detected cluster (Fig. 7b). The described co-expression analysis is available in Perseus through the R language interface provided by PluginInterop which interfaces directly with the established WGCNA library[42].

We applied the WGCNA co-expression analysis to parts of a cancer proteomics dataset[43], following the recommended workflow[7] from within Perseus. Bi-weight midcorrelation, a robust alternative to Pearson

[7] http://www.peterlangfelder.com/wgcna-resources-on-the-web/

correlation, was chosen to calculate correlations between all pairs of proteins. In order to obtain a scale-free co-expression network, a power parameter of 10 was selected (Fig. 7c), leading to an approximately scale-free network with a scale-free fit index of 0.9. Hierarchical clustering of the co-expression network identified 30 modules (Fig. 7d). The representative expression profiles of each of the modules, as provided by the corresponding module eigengene, were correlated with the available clinical annotations. This high-level overview over the data was then visualized in a heatmap (Fig. 7e). Several modules showed high correlations with specific clinical annotations. The magenta module showed high correlation with the triple-negative subtype (TN) and was highly enriched for the 'interferon-gamma-mediated signaling pathway' GO category. The top module hub genes with kME > 0.8 were GBP1, TAP1, TAPBP, HLA-A, TAP2, STAT1, and EML4. The purple module showed high correlation with Stage III, but and in depth look at the co-expression clustering heatmap revealed the module to be dominated by a single patient, limiting the validity of the module.

SOFTWARE IMPLEMENTATION, DOWNLOAD AND MAINTENANCE

The Perseus network module PerseusNet is implemented in the C# programming language using Visual Studio 2017, like the whole Perseus software. PerseusNet is distributed with Perseus by default and can be downloaded from http://www.perseus-framework.org. The current version, which is described in this manuscript, is 1.6.2.3. The PluginInterop and PHOTON plugins are also included in the standard download. In the current release, it is recommended to use Windows as operating system, although Linux support is underway, realized in the same way as for the MaxQuant software[44,45], by ensuring Mono compatibility. A plugin API enables external programmers to extend the functionality of PerseusNet and Perseus in general, by programming their own workflow activities. Plugin extensions by the user community will be linked from the plugin store at http://www.coxdocs.org/doku.php?id=perseus:user:plugins:store upon request. Context-specific documentation is linked from each activity (Supplementary Fig. 3). Step-by-step guides for the integration of external tools, such as Python or R, that have to be installed and configured separately from the main Perseus software, are available online[8]. A help forum for Perseus and PerseusNet is available at https://groups.google.com/group/perseus-list. Bugs that are reproducible in the latest available software version should be reported at https://maxquant.myjetbrains.com/youtrack.

DISCUSSION

We introduced PerseusNet, the network analysis extension for the Perseus software. It enables proteomics researchers to perform most network analysis by themselves. PerseusNet is highly extensible through a plugin API and its extension to R and Python, which allows for the incorporation of a plethora of existing scripts and programs from the network community. We envision that large part of the future programming will not be done by local developers, but by the global community through the plugin API. Programmers can release their plugins under licenses of their choice.

We have implemented powerful proteomics-specific activities for AP-MS network generation and PTM-related network analysis, presumably the two main applications for networking in proteomics. We plan to extend PerseusNet in the near future by activities from other proteomics sub-domains, as interaction determination by protein correlation profiling[46] and large scale network generation from cross-linking experiments on whole cell lysates[47].

ONLINE METHODS

CREATING INTERACTION NETWORKS FROM PULL-DOWN EXPERIMENTS

We created an interaction network from a pull-down screen[27] (Fig. 4). First, .RAW files were obtained from PRIDE (PXD003758) and processed with MaxQuant version 1.6.2.10. Mouse protein sequences were downloaded from UniProt (release 2017_07). Parameters 'matching between runs' and 'LFQ' were selected in addition to the default parameters. Downstream analysis of the 'proteinGroups.txt' output table was performed in Perseus. Columns for baits Eed, Ring1b, and Bap1 and their controls in the ESC and NPC cell lines were selected and log transformed. Quantitative profiles were filtered for missing values, and were filtered independently for each of the bait control pairs, retaining only proteins that were quantified in all three replicates of either the bait, or control, pull down. Missing values were imputed (width 0.3, down shift 1.8) before combining the tables,

[8] https://github.com/jdrudolph/PluginInterop

resulting in a total list of 2995 proteins on which the multi-volcano analysis was performed (Supplementary Table 1). The s0 and FDR parameters for Class A (s0=1, FDR=0.01%) and Class B (s0=1, FDR=0.2%) were chosen by visual inspection, aiming for a low number of significantly depleted proteins in any of the experiments. The interaction network was created by connecting significantly enriched prey proteins to their baits. Edges representing known protein complex interactions were annotated in the network. Due to missing mouse CORUM annotations for any of the baits, mouse CORUM annotations were obtained by mapping between mouse and human homologues as listed in the MGI[48] database. The resulting annotated network was then visualized in Perseus.

Approximately scale-free topology of the STRING interaction network

In order to investigate the topology of a large scale interaction network (Fig. 5), we first downloaded the human STRING interaction network (v10.5) from the website. After filtering for high confidence interactions (>0.9), the scale-free fit index was calculated according to reference[42]. Node degrees were calculated and plotted against their frequency distribution on a log-log scale. The R^2 of a linear fit to the log-log space represents the scale-free fit index.

NETWORK ANALYSIS OF A PHOSPHO-PROTEOMIC DATASET OF EGF STIMULATION

Two separate analyses, PHOTON and KSEA, were performed on the same experimental dataset[33] (Supplementary Table 2) and compared (Fig. 6). Log2 fold-changes for EGF from two replicates were averaged. For PHOTON analysis, we first generated a high-confidence PPI network. We downloaded all interactions from HIPPIE and filtered them for high confidence interactions (confidence > 0.72) and additionally removing high-degree nodes (degree < 700). The experimental data was mapped from UniProt to Entrez GeneIDs and subsequently used to annotate the nodes of the network. We then performed PHOTON analysis with adjusted default parameters. Network reconstruction with ANAT was enabled with the 100 highest scoring proteins and EGF anchor (GeneID 1950). Additionally, we increased the number of permutations to 100,000. The KSEA analysis was performed on the human site-specific kinase-substrate network from PhosphositePlus[17]. Data and network were matched based on UniProt identifiers. The resulting KSEA Z-scores and PHOTON signaling functionality scores were plotted against each other in Perseus. Proteins annotated with the GO category 'Epidermal growth factor receptor signaling pathway' (GO:0007173), were highlighted in red.

Co-expression analysis of a clinical proteomics dataset

We applied the co-expression network analysis workflow to a clinical proteomics dataset (Fig. 7). Protein quantification data and clinical annotation were obtained from Yanovic et al.[43]. SILAC ratios were first transformed to log(light/heavy). The dataset was filtered for the 43 patients unique to Yanovich et al. Using global hierarchical clustering of the patients, 4 outlier samples were identified and removed from the dataset. Additionally, proteins with less than 70% valid values were removed from the dataset and the resulting patient profiles were Z-scored (Supplementary Table 3). The power parameter for the co-expression analysis was selected using the 'Soft-threshold' activity. Using a signed network and the biweight midcorrelation, the power 10 was the lowest to a have scale-free fit index of more than 0.9 (Figure 7c). Proteins where then subjected to co-expression clustering (Figure 7d) and 30 co-expression modules were identified. The eigengene of each co-expression module was correlated with the provided clinical data using Pearson correlation and clustered using hierarchical clustering (Figure 7e).

PLUGININTEROP PROVIDES A CENTRAL ENTRY-POINT FOR ALL EXTERNAL PLUGINS

The PluginInterop project is written in the C# programming language and implements several Perseus plugin APIs. For users it provides a number of activities in Perseus for executing script files written in the Python and R languages. Upon selection of any of these activities, users will be prompted with a parameter window, allowing them to pass additional arguments to the script and requiring them to specify the executable that should be used for processing. Since Perseus does not include an installation of Python or R, users will have to install those and any other dependencies separately. PluginInterop aids the user by trying to automatically detect an existing installation and provide meaningful error messages in case of missing dependencies. Developers can additionally leverage the functionality implemented in PluginInterop as a basis for parametrized scripts. In general, developers are free to choose which external scripting language or program they would like to utilize. We found the R and Python scripting languages to be most useful, which is why we provide two companion libraries `perseuspy` and `PerseusR` to be used alongside PluginInterop. These libraries aid the communication between Perseus and the script.

The communication between Perseus and external scripts is straightforward and is easily implemented for

any tools of choice. In short, Perseus will persist all necessary data to the hard-drive and call the specified tool with specific command-line arguments. The first arguments contain all the parameters specified by the user, per choice of the developer, either in an XML format, or simply separated by spaces. Secondly, the input data from the workflow is saved to a temporary location which is passed to the script. The final arguments specify the expected location of the output data. The external process can provide status and progress updates to the user, as well as detailed error reporting by printing to stdout/stderr and indicating success or failure through the exit code. Once the process exits, Perseus will parse the output data for its expected location and insert it to the workflow. Any step in the pipeline is customizable for advanced scenarios, such as custom data formats.

The PluginInterop binary is automatically included in the latest Perseus version. The source code was published under the permissive, open-source MIT license, on Github[9]. The website also provides more information on how to develop plugins, including a video demonstration. The plugins presented in this manuscript are all developed on top of PluginInterop and the perseuspy and PerseusR companion libraries.

LIBRARY SUPPORT FOR SCRIPTING LANGUAGES

We implemented libraries in R and Python, which facilitate the interoperability of Perseus with external scripting languages. The main aim of these libraries is to map the data structures of Perseus to a counterpart native to the external language. Developers proficient in these languages will be more comfortable and productive with these native data structures. The largest benefit comes from the resulting integration with the existing data science ecosystem, all now available to Perseus plugin developers.
The `perseuspy` module provides data mappings for the Python language. The Perseus expression matrix is mapped the `DataFrame` object of the popular `pandas` module, which is tightly integrated with `numpy`, the de-facto standard for numerical computations in Python. The Perseus network collection data-type maps to a list of networks from the `networkx` package. It features a variety of graph algorithms and interfaces well with other modules, due to its usage of standard Python dictionaries. `perseuspy` is distributed via The Python Package Index (PyPI), allowing for easy installation of the module for developers and users alike. The code of `persuspy` is published under the permissive, open-source MIT license, and available alongside usage examples and more information on https://github.com/jdrudolph/perseuspy.
For the R language, we implemented the `PerseusR` package. It provides a mapping of the Perseus expression matrix to a custom wrapper class around the R `data.frame` object. The wrapping was necessary to represent Perseus-specific information such as annotation rows. Alternatively, developers can load data as a Bioconductor `expressionSet` object which enables the interface with the entire Bioconductors bioinformatics suite. Currently there is no support for network collections in `PerseusR`, but we plan to implement it in the near future. `PerseusR` is also published under the MIT license and its code is available on https://github.com/jdrudolph/PerseusR. Currently `PerseusR` is easily installed directly from the website. Due to the lengthy submission process, `PerseusR` will be uploaded to CRAN at a later point in time.

IMPLEMENTATION OF PLUGINPHOTON

We implemented a Perseus plugin for the PHOTON tool on top of the functionality provided by PluginInterop and perseuspy. PHOTON was previously capable to run only a single experiment at a time with a fixed human protein-protein interaction network. We expanded its implementation to allow for parallel processing of any number of experiments on any network. These changes make large datasets from any species directly amenable to PHOTON analysis. PluginPHOTON is published under the MIT licence, its code is available on https://github.com/jdrudolph/photon, and it is included in the latest Perseus release.

IMPLEMENTATION OF PLUGINCOEXPRESSION

We implemented parts of the WGCNA pipeline as a Perseus plugin. PluginCoExpression provides access to the WGCNA functions implemented in the R language via PluginInterop and PerseusR.

Implementation of KSEA in Perseus

KSEA analysis was implemented in Perseus and tested for correctness against the reference implementation.

ACKNOWLEDGEMENTS

We thank J. Sebastian Paez and Sung-Huan Yu for contributing to PerseusR, and Caroline Friedel and Tamar Geiger for helpful discussions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 686547 and from the FP7 grant agreement GA ERC-2012-SyG_318987–ToPAG.

[9] https://github.com/jdrudolph/PluginInterop

AUTHOR CONTRIBUTIONS

J.R. and J.C. planned and performed the research, developed the software and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Bar-Yam, Y. General Features of Complex Systems. *Knowl. Manag. Organ. Intell. Learn. Complex.* **I,** 1–10 (1997).
2. O'Connor, T. & Wong, H. Y. Emergent Properties. *Stanford Encycl. Philos.* 1–25 (2012). doi:10.1111/1467-9973.00225
3. Grandjean, M. A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts Humanit.* **3,** (2016).
4. Goffman, C. And What is Your Erdös Number? *Am. Math. Mon.* **76,** 791 (1969).
5. Sporns, O., Tononi, G. & Kötter, R. The human connectome: A structural description of the human brain. *PLoS Computational Biology* **1,** 0245–0251 (2005).
6. Seung, S. *Sebastian Seung: I am my connectome | Talk Video | TED.com*. *Ted.com* (2010).
7. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5,** 101–113 (2004).
8. BUTTE, A. J. & KOHANE, I. S. MUTUAL INFORMATION RELEVANCE NETWORKS: FUNCTIONAL GENOMIC CLUSTERING USING PAIRWISE ENTROPY MEASUREMENTS. in *Biocomputing 2000* 418–429 (1999). doi:10.1142/9789814447331_0040
9. 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).
10. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13,** 731–40 (2016).
11. Tyanova, S. & Cox, J. in *Methods in Molecular Biology* **1711,** 133–148 (2018).
12. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
13. Szklarczyk, D. *et al.* The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
14. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45,** D369–D379 (2017).
15. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gkt1115
16. Ruepp, A. *et al.* CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res.* **38,** (2009).
17. Hornbeck, P. V *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43,** D512-20 (2015).
18. Gingras, A. C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol* **8,** 645–654 (2007).
19. Dunham, W. H., Mullin, M. & Gingras, A. C. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics* (2012). doi:10.1002/pmic.201100523
20. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163,** 712–723 (2015).
21. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162,** 425–440 (2015).
22. Hubner, N. C. *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J Cell Biol* **189,** 739–754 (2010).
23. Noble, W. S. How does multiple testing correction work? *Nature Biotechnology* **27,** 1135–1137 (2009).
24. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98,** 5116–5121 (2001).
25. Franz, M. *et al.* Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics* **32,** 309–311 (2015).
26. Dogrusoz, U., Giral, E., Cetintas, A., Civril, A. & Demir, E. A layout algorithm for undirected compound graphs. *Inf. Sci. (Ny).* **179,** 980–994 (2009).
27. Kloet, S. L. *et al.* The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nat. Struct. Mol. Biol.*

23, 682–690 (2016).
28. Pedamallu, C. S. & Ozdamar, L. A Review on protein-protein interaction network databases. in *Springer Proceedings in Mathematics and Statistics* **73,** 511–519 (2014).
29. Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw985
30. Gene Ontology, C. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43,** D1049-56 (2015).
31. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* (2009). doi:10.1137/070710111
32. Riley, N. M. & Coon, J. J. Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. *Anal. Chem.* **88,** 74–94 (2016).
33. Jan Daniel Rudolph, Marjo de Graauw, Bob van de Water, Tamar Geiger & Roded Sharan. Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. *Cell Syst.* **3,** 585–593 (2016).
34. Casado, P. *et al.* Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* **6,** rs6-rs6 (2013).
35. Hernandez-Armenta, C., Ochoa, D., Gonçalves, E., Saez-Rodriguez, J. & Beltrao, P. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* **33,** 1845–1851 (2017).
36. Herranz, N. *et al.* mTOR regulates MAPKAPK2 translation to control the senescence-associated secretory phenotype. *Nat. Cell Biol.* (2015). doi:10.1038/ncb3225
37. Wilkes, E. H., Terfve, C., Gribben, J. G., Saez-Rodriguez, J. & Cutillas, P. R. Empirical inference of circuitry and plasticity in a kinase signaling network. *Proc. Natl. Acad. Sci.* (2015). doi:10.1073/pnas.1423344112
38. Linding, R. *et al.* NetworKIN: A resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* **36,** (2008).
39. Wiredja, D. D., Koyutürk, M. & Chance, M. R. The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx415
40. Noya, S. B. *et al.* Rest-activity cycles drive dynamics of phosphorylation in cortical synapses. *submitted*
41. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4,** Article17 (2005).
42. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
43. Yanovich, G. *et al.* Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification. *Cancer Res.* (2018).
44. 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).
45. Sinitcyn, P. *et al.* MaxQuant goes Linux. *Nat Methods* (2018).
46. Kristensen, A. R. & Foster, L. J. Protein correlation profiling-SILAC to study protein-protein interactions. *Methods Mol. Biol.* (2014). doi:10.1007/978-1-4939-1142-4_18
47. Liu, F., Lössl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* **8,** 15473 (2017).
48. Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E. & Bult, C. J. Mouse Genome Database (MGD)-2018: Knowledgebase for the laboratory mouse. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1006

FIGURES

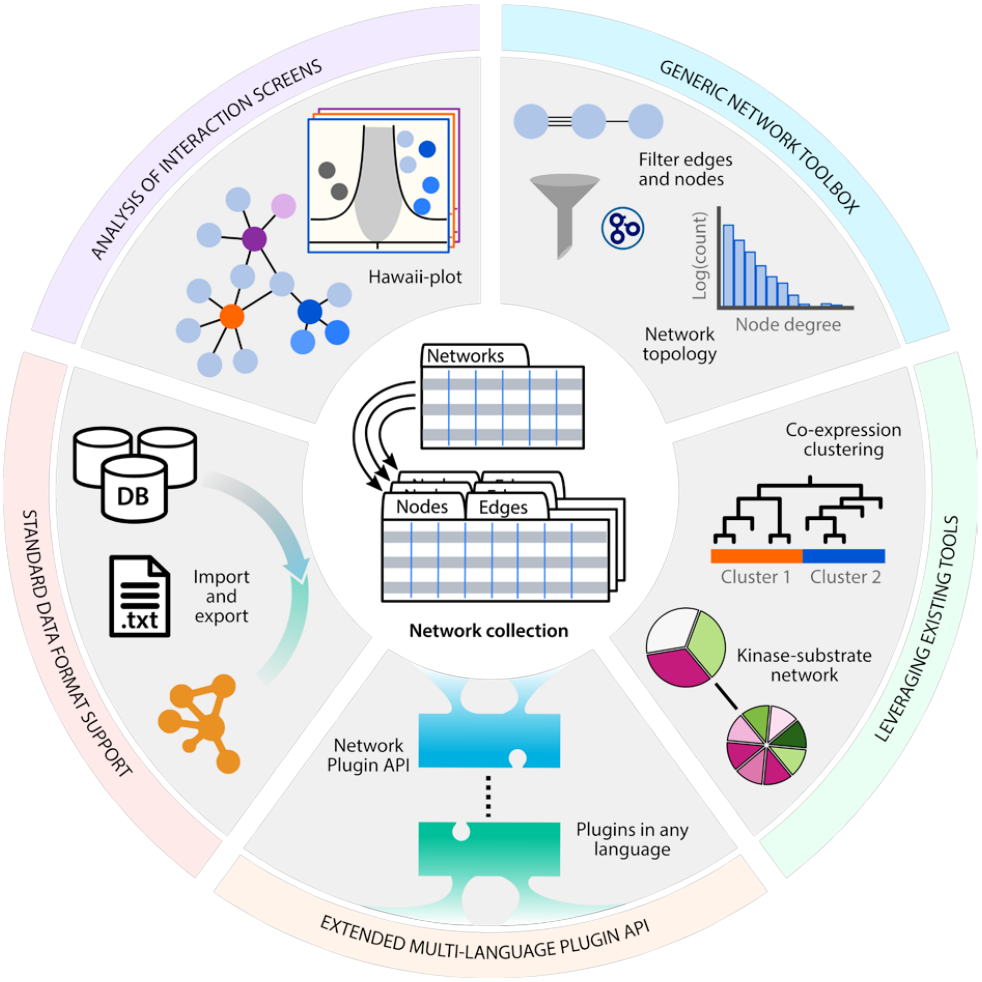


Figure 1 Schematic overview of the new network functionality in Perseus. PerseusNet implements a number of processing and analysis steps facilitated by the network collection data type. While including proteomics centric analyses, such as for the analysis of interaction screens, the network module also provides a number of general purpose tools, as, for instance, for network annotation, filtering, and topology determination. With the extension of the Perseus plugin API to networks and furthermore to other programing languages, it becomes possible to integrate existing network analysis tools in Perseus. Networks are easily imported to, and exported from Perseus, due to its support for standard formats.

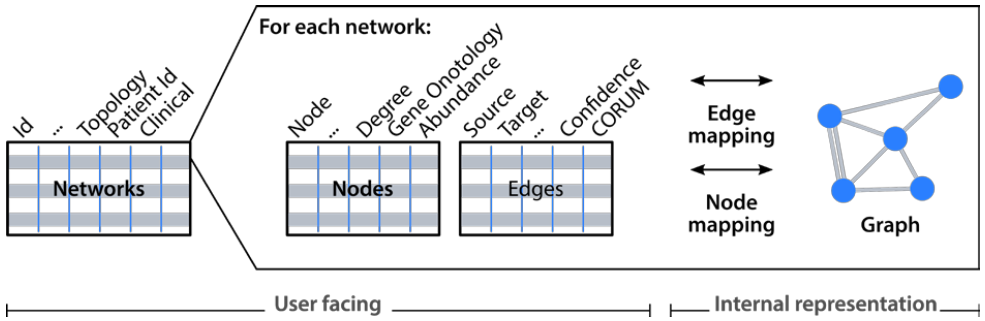


Figure 2 Schematic representation of the network collection data type. User facing information is displayed in tabular form with tables listing the networks in the collection, as well as providing detailed information on the nodes and edges of each network. Internally an auxiliary graph data structure aids in the implementation of graph algorithms. Node- and edge-mapping provide the required cross-references between the tabular and graph representation.

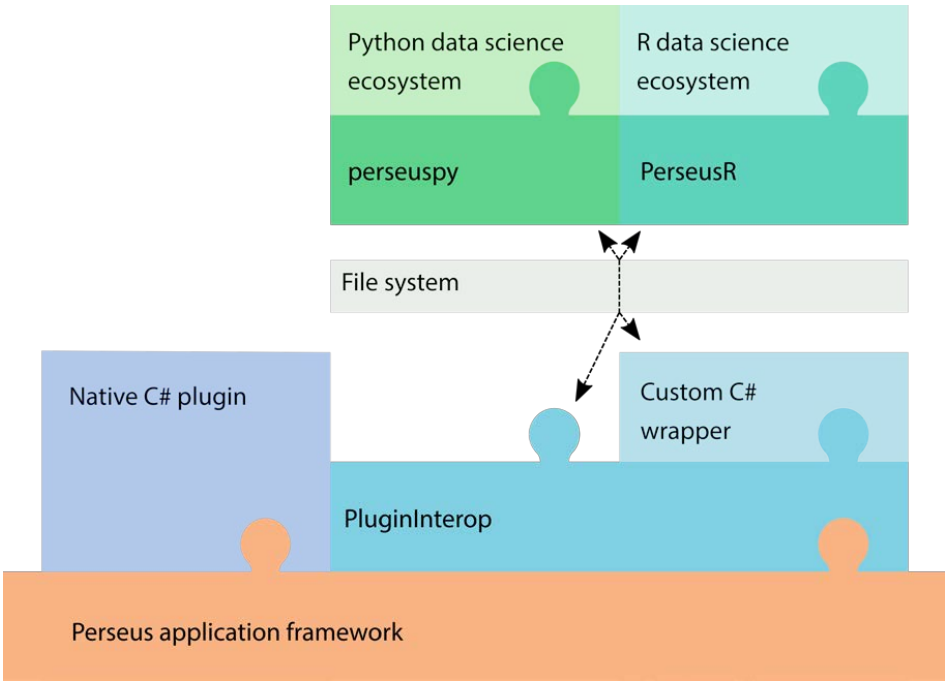


Figure 3 Schematic of the Perseus plugin system. Plugins written in C# are native to Perseus and implement their functionality directly on top of the APIs and data structures provided by the application framework. PluginInterop enables the execution of scripts in the Python and R languages, as well as other external programs. By communicating via the file system, data is transferred between Perseus and the external program. The companion libraries perseuspy and PerseusR enable developers to access the data science ecosystem in their language of choice. For custom GUI elements and an improved user experience of external tools, developers can implement a thin C# wrapper class that extends the generic functionality of PluginInterop.

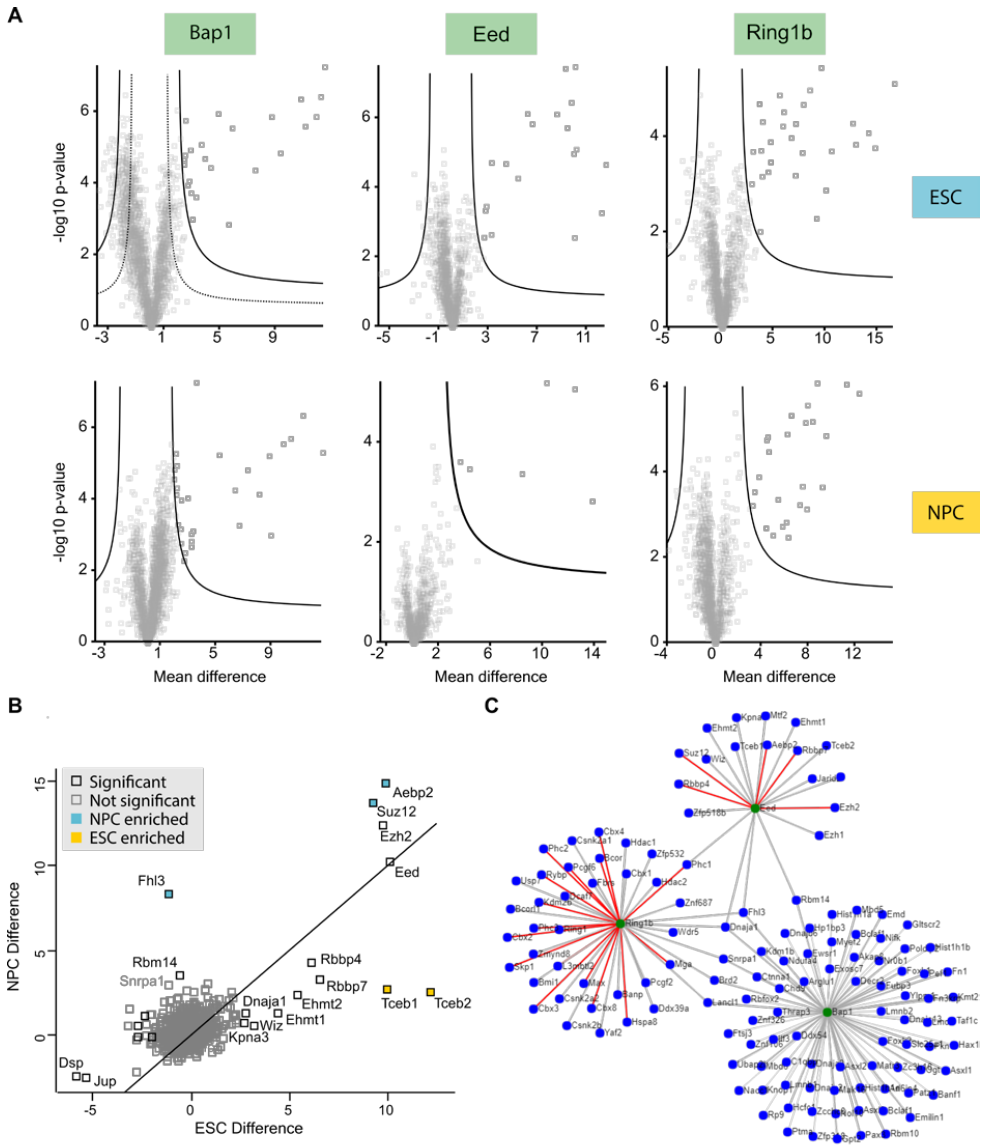


Figure 4 AP-MS. **a** The Hawaii plot provides an overview over the entire dataset[27] (Supplementary Table 1 and online methods) consisting of three baits in two conditions. Significant interactors are determined using a permutation-based FDR and the resulting Class A (solid line) and Class B (dashed lines) thresholds are displayed in the plot. Class A interactors are displayed in dark grey, other proteins are shown in light grey. **b** Enrichment plot comparing the Eed pull downs in ESC and NPC cell lines. Significant interactors in any of the two conditions are displayed in black, non-significant proteins are displayed in light grey. Proteins differentially enriched in one of the two conditions will be located far from the diagonal and can be identified visually. **c** Visualization of the resulting protein interaction network for both cell lines. Bait proteins are colored in green and their interactors are colored in blue. Thick lines represent Class A interactions, thinner lines Class B. Interactions which were already annotated in the human CORUM database are highlighted in red.

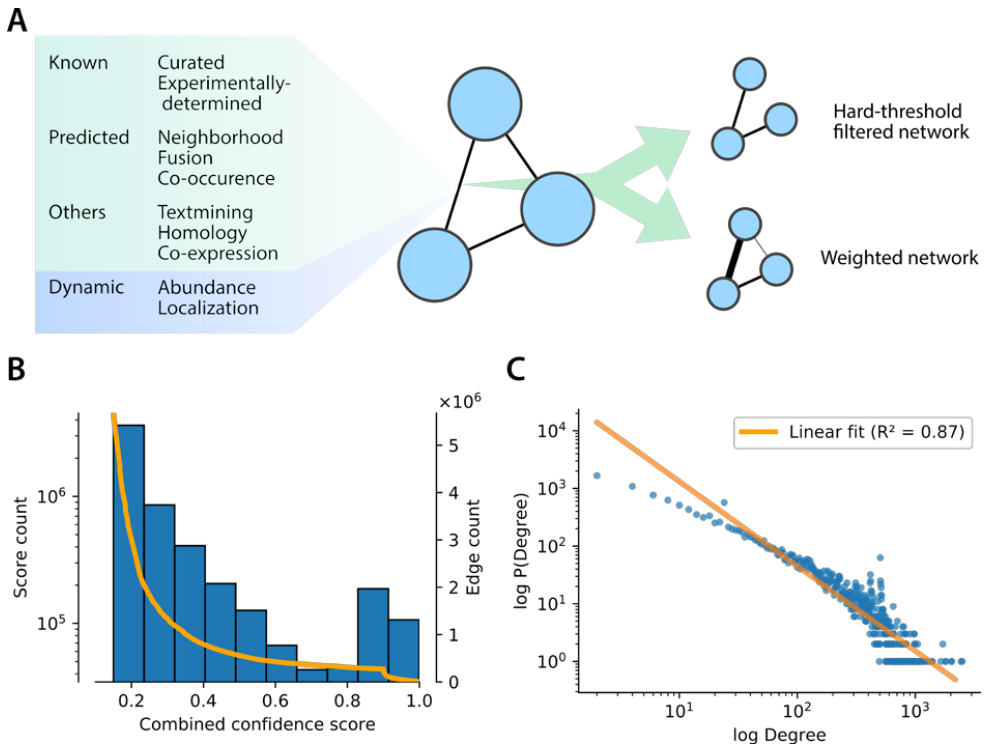


Figure 5 Handling large-scale protein interaction databases in Perseus. **a** Interactions in PPI databases are often annotated with confidence scores derived from various sources. Static confidence scores can combine experimental evidence and predictions of physical interaction as well as non-physical interactions between the proteins. Confidences can be adjusted dynamically based on condition-specific data to better represent the changed wiring. In any analysis a high-confidence network can be obtained by removing edges below a given hard threshold. Many analysis can directly utilize confidence scores as so-called edge weights, thereby allowing for the inclusion of lower-confidence interactions. **b** Histogram of the combined confidence score from the human STRING PPI network. Superimposed in orange is the number of interactions in the filtered network if the edges with scores lower than the current value were removed. Filtering out low confidence edges leads to a significant reduction in the number of edges in the final network. **c** Log-log plot of the node degree against the degree frequency generated from the human STRING PPI network. The R^2 value of the linear fit (orange) to these data represents the scale-free fit index.

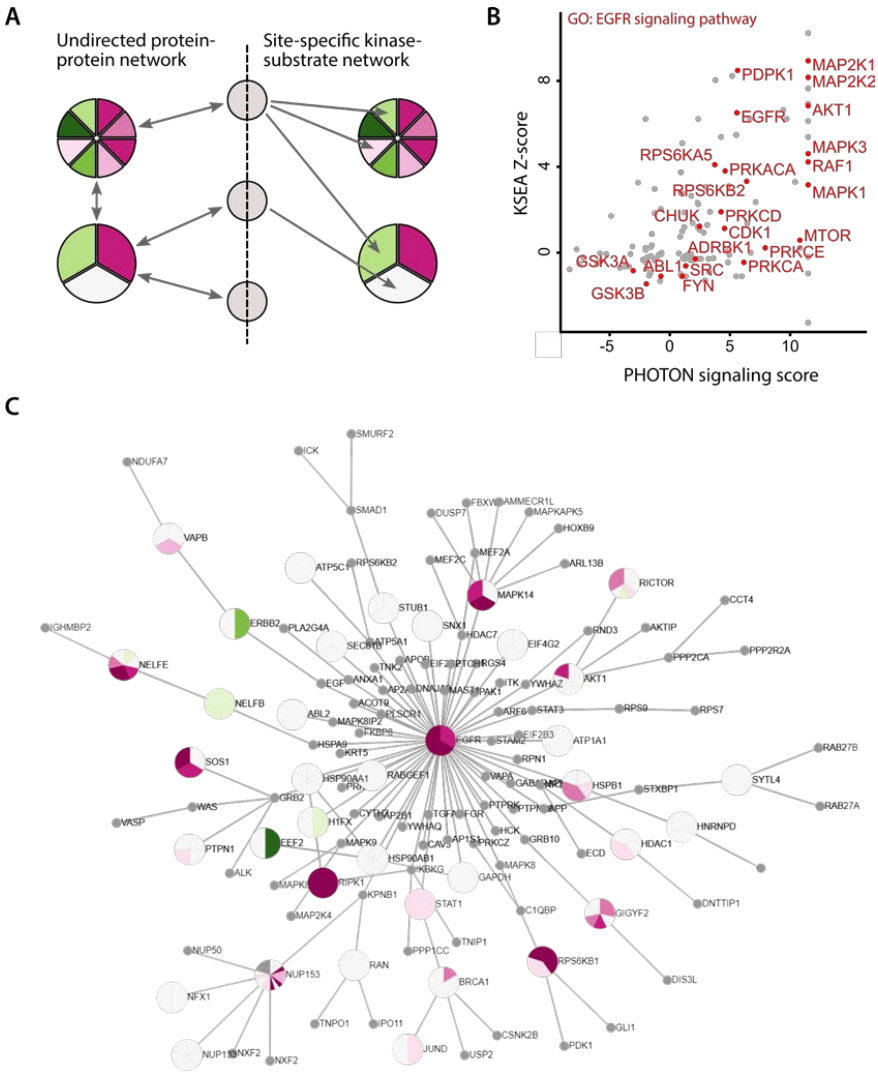


Figure 6 Network analysis of an EGF stimulation phospho-proteomics study. **a** Comparison of network topologies used for the analysis of phospho-proteomic data. Nodes in the network are represented as grey circles or pie-charts where each slice represents the observed phosphorylation changes at a specific site on the protein. Physical protein-protein interactions (left side) are present between all classes of proteins and are by definition undirected. In order to capture the enzymatic action of kinases more accurately, directed interactions (right side) from kinase to substrate are defined in a site-specific manner. **b** KSEA Z-score and PHOTON signaling functionality scores derived from phospho-proteomic data measured after EGF stimulation (Supplementary Table 2) only weakly correlate to each other (Pearson correlation 0.52). Kinases annotated in GO with the term 'Epidermal growth factor receptor signaling pathway` are highlighted in red. Both methods assign high scores to central members of the expected pathway. **c** Signaling network reconstructed by PHOTON from the 100 highest scoring proteins anchored at EGF. The interactive visualization has an automatic layout and phosphorylation data overlay.

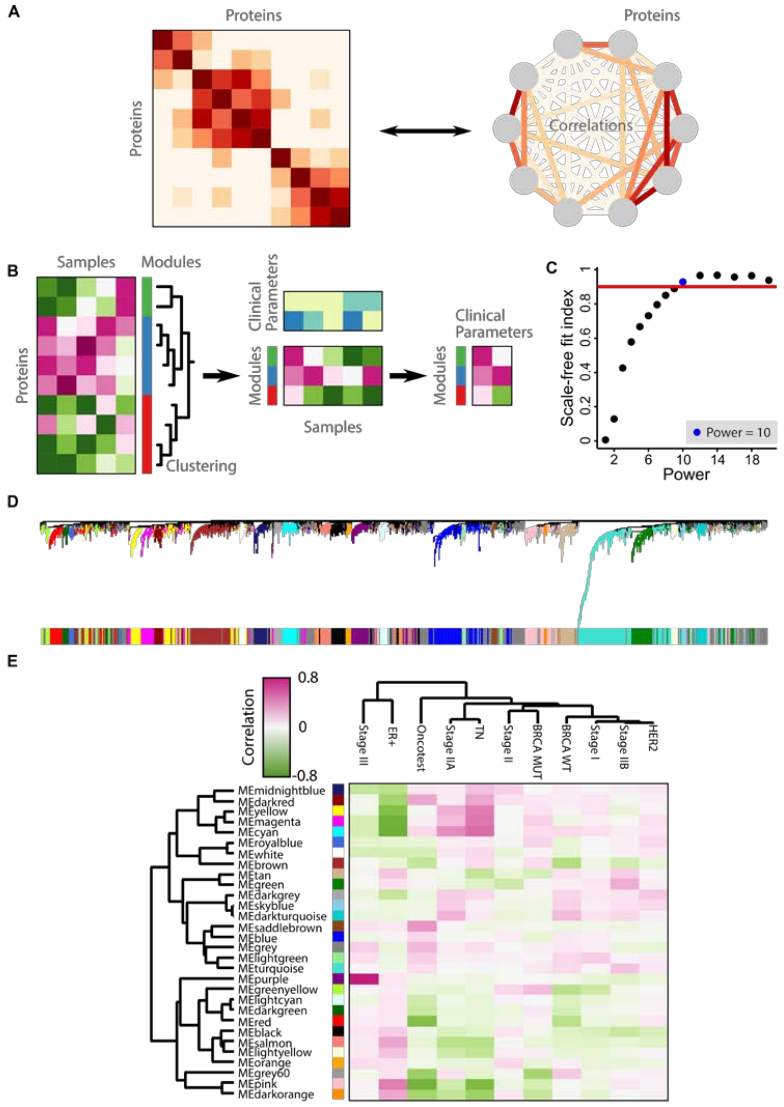


Figure 7 Co-expression network analysis on clinical data. **a** Any correlation matrix can be interpreted as a fully-connected network with edge weights corresponding to the correlation between the proteins. Hierarchical clustering of the correlation matrix can utilize a network-based distance function. **b** Co-expression clustering and identified co-expression modules annotate the original expression matrix. Phenotype data can be correlated with representative co-expression module profiles and provide a high-level interpretation of the modules. **c** Parameter selection of the power parameter for the Yanovich et al.[43] dataset (Supplementary Table 3 and Online Methods). The lowest power reaching close to a high scale-free fit index of 0.9 (red line) was selected. **d** Co-expression cluster dendrogram and assigned modules. **e** Correlation heat map between module eigengenes and clinical parameters.

2.2 Phosphoproteomics of cortical synapses

The following manuscript is the second in a series of two manuscripts studying the effect of the circadian clock and sleep wake pressure on the transcriptome, proteome and phosphoproteome of cortical synapses. It is the first report on the circadian regulation of cortical synapses, which were previously studied only at the lower temporal resolution of sleep and wake states [Wang et al., 2018, Diering et al., 2017].

In the first manuscript, the transcriptome and proteome of forebrain synaptosomes was analyzed at three different times of day with state-of-the art technology. 70% of transcripts were found to exhibit daily dynamics that correspond to the light to dark and dark to light transitions. The transcripts peaking at each transition were found to relate to synaptic function during the light to dark transition, and metabolism and translation at the dark to light transition. The proteome data corroborated these findings under wild-type conditions. Under sleep pressure, unlike the transcripts, none of the detected proteins retained their daily dynamics. In conclusion, the circadian clock was found to drive transcription, while protein synthesis was driven by the actual sleep to wake transitions.

The second manuscript focusses on the phosphoproteomic characterization of the daily dynamics of synaptosomes. A periodicity analysis of the more than 14,000 identified phosphopeptides found around 30% of the peptides to have an oscillatory behavior. Most phosphopeptides peaked during the light to dark and dark to light transitions and were found to be associated with synaptic transmission, cytoskeleton reorganization and excitatory/inhibitory balance. In comparison, after sleep deprivation, 95% of the phosphoproteome did not exhibit any time-of-day dependent dynamics. By analyzing the network of kinases and their substrates, we identified rhythmically active kinases which were themselves controlled by temporal phosphorylation.

My contribution to the study was the analysis of the circadian phosphoproteome by extending and applying the PHOTON algorithm [Rudolph et al., 2016]. The extension of PHOTON to handle multiple conditions and any PPI network [Rudolph and Cox, 2018] allowed me to apply it to this dataset in combination with the mouse STRING [Szklarczyk et al., 2017] interaction network, and calculate signaling functionality scores for all kinases. The results are presented in Figure 4 and described in the results and methods sections.

Sara B. Noya, Franziska Brüning, Tanja Bange, Jan D. Rudolph, Jürgen Cox, Steven A. Brown, Matthias Mann, and Maria S. Robles. Rest-activity cycles drive dynamics of phosphorylation in cortical synapses. *Submitted*, 2018

Rest-activity cycles drive dynamics of phosphorylation in cortical synapses

Sara B Noya[1]†, Franziska Brüning[2,3]†, Tanja Bange[3], Jan D Rudolph[4], Shiva Tyagarajan[1], Jürgen Cox[4], Matthias Mann[2], Steven A Brown[1]* and Maria S Robles[3]*§

[1] Institute of Pharmacology and Toxicology, UZH, Zurich, Switzerland
[2] Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried 82152, Germany
[3] Institute of Medical Psychology, LMU Munich, Germany
[4] Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

† contributed equally to this work

* Correspondence to: charo.robles@med.uni-muenchen.de, steven.brown@pharma.uzh.ch; § lead contact

Abstract

The circadian clock drives daily changes of physiology, including sleep-wake cycles, by regulating transcription, protein abundance and function. Circadian phosphorylation controls cellular processes in peripheral organs, but little is known about its role in brain function and synaptic activity. We applied advanced quantitative phosphoproteomics to mouse forebrain synaptoneurosomes isolated across 24h, accurately quantifying almost 8,000 phosphopeptides. Remarkably, half of the synaptic phosphoproteome, including numerous kinases, had large-amplitude rhythms peaking at rest-activity and activity-rest transitions. Bioinformatic analyses revealed global temporal control of synaptic function via phosphorylation, including synaptic transmission, cytoskeleton reorganization and excitatory/inhibitory balance. Remarkably, sleep deprivation abolished 98% of all phosphorylation cycles in synaptoneurosomes, indicating that rest-activity cycles rather than circadian signals are main drivers of synaptic phosphorylation, responding to both sleep and wake pressures.

One Sentence Summary

Sleep and wake pressure shape synaptic phosphorylation dynamics and function

Introduction

Circadian clocks are endogenous oscillators present in virtually every mammalian cell. The molecular mechanism of the clock drives cycles of transcription, translation and protein activity to regulate daily changes in physiology and behavior. Mass spectrometry (MS) - based quantitative proteomics has importantly contributed to our understanding how circadian post-transcriptional mechanisms temporally shape metabolic processes in peripheral tissues (*1, 2*). Interestingly, circadian phosphorylation changes by far eclipse the regulation at the transcriptional and proteome levels in amplitude (*3*). Temporal characterization of proteome and phosphoproteome changes in the central nervous system, in contrast, has been challenging due to the sensitivity, dynamic range and throughput required to capture the regional, cellular and synaptic heterogeneity. However, recent advances in MS in combination with spatial isolation methods allow the deep characterization of proteomes from different brain regions and cell populations (*4, 5*). In addition, high-throughput phosphoproteomic technologies are now suitable for the global characterization of phosphorylation signaling dynamics in different brain areas (*6*).

Numerous synaptic features, like diffusion of receptors in membranes, channel conductance or cytoskeleton remodeling, depend on fast phosphorylation-based control mechanisms. In particular, synaptic plasticity and scaling have been linked to phosphorylation of receptors, scaffolding, cytoskeletal and other synaptic proteins (*7, 8*). Although quantitative phosphoproteomics has been applied to the synaptic compartment, technical limitations have so far precluded accurate quantification that would allow the precise characterization of global phosphorylation dynamics associated with synaptic function (*7*). It is thus unknown whether daily changes in synaptic activity are coupled to global dynamics of phosphorylation in synapses, and moreover, whether daily rhythms of phosphorylation temporally segregate synaptic processes. Two recent reports have addressed these technical limitations by either fractionating post-synaptic density (*9*) or by mapping whole-brain phosphoproteomics to synaptic protein annotations (*10*). Both highlight a role for sleep pressure in driving synaptic phosphorylation changes associated with the kinase (SIK3) and downstream effectors (HOMER1a).

Here we apply state-of-the-art quantitative MS-based proteomics to characterize *in vivo* phosphorylation dynamics across the day in isolated synaptoneurosomes from mouse forebrain, resulting in by far the most comprehensive time-resolved phosphoproteome of synapses to date. Combination with in depth proteomics (see accompanying paper (*11*)) reveals that more than one fourth of the individual phosphorylation sites in synaptic proteins oscillate daily and independently of protein abundance. We bioinformatically analyze dynamic synaptic process and the regulatory protein classes modulated by phosphorylation.

Temporally modulated phosphorylation networks gate synaptic processes at both dawn and dusk, suggesting a potential association with activity-rest cycles. To test this, we interfere with the sleep cycle and investigate the effect on the synaptic phosphoproteome. This leads to a dramatic ablation of global phosphorylation cycles, suggesting a dominant role of both sleep and wake pressure in synaptic phosphorylation dynamics.

Results

In-depth phosphoproteomic profiling of synaptoneurosomes

To characterize daily dynamics of phosphorylation abundance specifically in synapses, we biochemically isolated synaptoneurosomes from mouse forebrains (*11*). Mice were kept in 12h:12h (light : dark) schedules and then sacrificed in biological quadruplicates at six time points, every 4 h, across 24 h (n=24 mice). We used a rapid method based on Percoll gradients to prepare synaptoneurosomes from forebrains, containing both pre- and post-synaptic components (*12*) and immediately flash-froze them to prevent de-phosphorylation. To achieve sufficient throughput for our time-dependent experiments, we employed the EasyPhos method (*13*) to enrich phosphopeptides from only 1 mg of proteome homogenate for each synaptoneurosome preparation. Figure 1A depicts the MS-based quantitative phosphoproteomics workflow consisting of single runs on a high-resolution, high sensitivity quadrupole-Orbitrap HF-X mass spectrometer. Across all samples this resulted in a total of 10,439 unique phosphosites in 14,462 phosphopeptides, mapping to more than 2,000 proteins (**Fig. 1B** and table S1). Comparing phosphorylated

amino acids in synapses to our previous circadian study in the liver (*3*) revealed similar proportions (83.6 % pS, 15.7% pT, 0.6% pY; Fig. 1B). Phosphopeptide intensities between measurements were highly reproducible in both biological replicates and time points (Pearson r = 0.88 and 0.83, respectively; **Fig. 1C**). Synaptic phosphopeptide intensities ranged over five orders of magnitude (fig. S1A), similar to what we found in liver, indicating that the synaptic compartment still has a wide quantitative range of phosphorylation levels. To indirectly evaluate our isolation method, we performed a Fisher's exact test on the total phosphoproteome dataset. Of all keywords, the top two are "Synapse" and "Cell junction" ($p < 10e-40$ for both) and the other highly enriched ones are also all relevant to synaptic function, even when analyzing every time point separately (**Fig. 1D** fig. S1B and S1C and see Methods). Our data show the power of combining high-throughput phosphoproteomics with biochemical isolation of synaptoneurosomes to deeply profile phosphorylation in this discrete neuronal compartment.

Daily rhythms of the synaptic phosphoproteome

Next we performed statistical cycling analysis in the circadian module of the Perseus software (*2, 3, 14*), to filter and cosine-fit the phosphopeptide intensities. A total of 2,202 (30.4%) of the 7,257 phosphopeptides accurately quantified in at least 50% of the samples oscillate in abundance with a rhythm of 24h (q-value < 0.05, see Methods, **Fig. 2 A and B** and table S2). We detected rhythmic phosphorylation for more than half of the synaptic phosphoproteins (838 out of the total 1,655, **Fig. 2C**). These rhythmic sites were localized with high probability to a single residue (mean 0.97) and their amino acid distribution was similar to the total dataset (**Fig. 1C** and fig. S2A). Cycling phosphopeptides had a similar intensity distribution as the total phosphoproteome (fig. S2B), implying that circadian phospho-regulation is not biased by abundance. Little is known about the magnitude of dynamic phosphorylation changes in the synaptic compartment and our data now revealed that these changes are substantial: the mean amplitude changes are more than three-fold, with hundreds of sites at more than 10-fold (**Fig. 2D**).

To assess to which extent these phosphorylation dynamics depend on protein abundance changes, we quantitatively compared the levels of phosphopeptides with the abundance of the corresponding protein. Almost 90% of proteins with cycling phosphopeptides were quantified at the protein level in our companion study (*11*), and of these only 5% significantly oscillate in abundance ($q < 0.05$, 24h period, **Fig. 2E**). Even the small fraction of rhythmic proteins with oscillating phosphorylation generally displays different phases across the day, and furthermore, multiple sites in the same protein generally behaved differently (**Fig. 2F**, fig. S2C). In the minor population of rhythmic proteins carrying cycling phosphorylation, the mean amplitudes at the phosphorylation level were ten-fold larger than those at the protein level (**Fig. 2G**). Our data clearly establish that protein phosphorylation in forebrain synaptoneurosomes is highly dynamic across the day and almost completely independent of protein abundance, suggesting another layer of synaptic functional regulation.

Temporal compartmentalization of synaptic protein phosphorylation

Our previous study in liver revealed that dynamic phosphorylation drives daily organ functions to an previously unappreciated degree (*3*). Examining the phosphorylation rhythms in the synaptic compartment showed that the phases of rhythmic phosphopeptides gathered in two distinct clusters. The larger one, at the light to dark transition when mice start to be active, contains two thirds of them, whereas the remaining largely clusters at the end of the night, preceding the resting phase (**Fig. 3A**). This phase distribution indicates a major rewiring of protein phosphorylation and presumably synaptic function at the activity-to-rest and rest-to-activity transitions. In order to identify synaptic functions that are temporally compartmentalized by protein phosphorylation, we searched for statistically enriched protein annotations in each of the two defined phase clusters (Fisher's exact test $p<0.05$, see Methods). At the end of the resting phase, our analysis found keywords representing Cell adhesion and Cell junction as well as Ion channels, Ion transporters, Hydrolases and Kinases highly enriched. In contrast, Cell projection, Cytoskeleton and Ubiquitin conjugation proteins are rhythmically phosphorylated in the cluster at the end of the activity phase (**Fig. 3, B and C**, table S3). This highlights the temporal regulatory nodes involved in the remodeling

processes that are known to occur at synapses (*15*). Proteins involved in Cell division and Mitosis were also overrepresented in the phosphorylation cluster at the end of the activity phase. Although far from the nucleus, the localization and function of some proteins with key roles in cell division, e.g. CDK5, are also important for synaptic activity (*16*), and such unexpected connections may be true for other proteins. We suggest that the combination of an in-depth, quantitative workflow with spatial isolation can enable the identification of phosphorylation events in proteins not traditionally associated with synaptic function.

Synapses are hubs of kinases

We next focused our attention on the major and specific enrichment of kinases, key regulators of almost all cellular process, at the end of the resting phase. Almost 500 phosphorylated peptides from a total of 128 kinases from all major families. Thus a fifth of the total mouse kinome is not only present in the synaptic compartment but also detectable in a phosphorylated form (fig. S3A). More than half of these kinases show at least one rhythmic phosphorylation with the same overall phase distribution as the total cycling synaptic phosphoproteome ($q < 0.05$, **Fig. 4A**, fig. S3B, 3C and table S4). They belong to all major kinase families with a higher representation of AGC threonine/serine kinases (**Fig. 4B** and fig. S3D). All of the 66 kinases with rhythmic phosphorylation were also quantified at the protein level in this compartment (*11*), however, only four of them cycled in protein abundance (fig S3E). Therefore, phosphorylation, rather than protein abundance, regulates temporal kinase function at synapses across the day.

Since site-specific phosphorylation does not always imply changes in activity, we set out to identify temporally activated kinases with an unbiased workflow that uses high-confidence protein-protein interaction networks and large-scale phosphorylation data to retrieve protein signaling functionality. This PHOTON pipeline assigns a score to each protein based on the phosphorylation status of their interacting proteins (*17*). As these scores reflect the changes in phosphorylation levels of their substrates/interactors, kinases that are activated rather than only phosphorylated should have PHOTON scores cycling across the 24h, with the maximum score indicating the peak of kinase activity. From all synaptic kinases with rhythmic phosphorylation, this analysis resulted in rhythmic activity patterns for 13 of them (see Methods). Of these,

11 are active at the rest-activity transition, including protein kinase C (PRKCA, PRKCB, PRKCG) and Ca2+/calmodulin-dependent kinase 2 (CAMK2B, CAMK2G). Conversely, the tyrosine-protein kinase ABL2 and the serine/threonine-protein kinase DCLK1 showed the opposite behavior, peaking at the activity-rest transition (**Fig 4, C and D**). Interestingly, the kinases activated at the end of the rest phase are associated with excitatory synaptic activity (*8*) while those activated at the end of the activity phase are associated with inhibitory synaptic activity (*18, 19*). PHOTON scores were also used to predict temporally regulated synaptic processes by means of phosphorylation dynamics (see Methods). In line with our kinase activity results, we inferred that the triggering of inhibitory synaptic mechanism, involving GABA, happens at the end of the activity period, while glutamate-mediated synaptic excitatory activity was predominantly associated to the rest-activity transition (**Fig. 4E**). These data are consistent with the roles of these synaptic types in rest and activity, respectively (*20, 21*). While the existing network data are sufficient to associate rhythmic activity with a substantial subset of the kinases, more complete networks will likely establish activity changes for a much larger fraction of the cycling synaptic phosphoproteome in the future. Together the combination of predictive and experimental data places extensive temporal regulation of kinases activity at the core phospho-dependent functional processes at the synapse.

Sleep deprivation abrogates synaptic phosphorylation rhythms

We hypothesized that the sharp distribution of synaptic phosphorylation patterns concomitant with the activity transitions may reflect sleep pressure (a sleep homeostat) in addition or alternatively to circadian (time-of-day) mechanisms. To test their relative contribution, we subjected mice to 4h sleep deprivation (SD) by gentle handling (*22*) prior to each timepoint, and collected brains every 6h in 24h (n=4/time point; **Fig. 5A**). This protocol of SD across the 24h time course equalizes the sleep pressure to keep it constantly high (*23*). Synaptoneurosomes from forebrain were prepared from SD mice, and phosphopeptides enriched prior to MS analysis as above (**Fig. 5A**, see Methods). We accurately quantified 7,021 phosphopeptides in at least 50% of the measured samples, very similar values to those in the base-line experiment, with more than 90% overlap between them (fig. S4A). Strikingly, cycling analysis of the 6,526 phosphopeptides

revealed almost complete abrogation of rhythmic phosphorylation in the synaptic compartment of SD mice. Only 2.3% (47 phosphopeptides corresponding to 41 proteins) cycled in SD (period=24h, $q < 0.05$, see Methods, **Fig. 5, B and C**, fig. S4B-D, table S5). These few remaining cycling phosphopeptides oscillated with similar amplitudes and with comparable phases in both conditions (**Fig. 6A-C**).

Since sleep pressure did not affect the pattern of the remaining rhythmic phosphopeptides, they appear to only be driven by the circadian clock. Of the corresponding 41 synaptic phosphoproteins, 31 form a protein interaction network with SHANK3, RTN4, MTAP1B, STX1, MYO5A and HSP90 at the core, much more than expected by chance (see Methods, fig. S5). They belong to interconnected cellular structures such as cytoskeleton, synaptic scaffolding, membrane, vesicle trafficking, and ubiquitin mechanisms, all important for synaptic integrity and function (*24-26*). Almost a third are cytoskeletal proteins with molecular motors such as MTAP1B and MYO5A and other microtubule associated proteins (MTAP4, KIF21A, STMN3, STMN1, ABI1). Furthermore, under SD rhythmic phosphorylation is also preserved on proteins involved in synaptic vesicle formation and exocytosis, which is essential for chemical neurotransmission (CHGB, STX1A, STXBP5). Cell adhesion (RTN4, PGRMC1, BASP1, NRCAM, CADM1) and scaffolding proteins (SHANK3, PICCOLO, BASSOON) are likewise unaffected, commensurate with the crucial role of these proteins in the regulation of synaptic integrity. Finally, the rhythmic phosphorylation of chaperone HSP90A, important in the synapse for trafficking of AMPA receptors (*24*), persisted. Thus only a very small number of phosphorylation sites in key synaptic components remain cycling under high sleep pressure, perhaps because of indispensable regulatory roles. Our results clearly establish that sleep deprivation severely affects synaptic phosphorylation across the day, leading to a dramatic loss in rhythmicity.

Discussion

Synaptic plasticity and function dynamically change across the day (*27*). It was already known that changes in synaptic activity are associated with the phosphorylation of several signaling proteins (*8, 26*). However, our large-scale quantitative phosphoproteomes of isolated synaptoneurosomes resulted in the first

comprehensive time-resolved map of synaptic phosphorylation across the entire activity and rest phases. We detected cycles on 30% of phosphorylation events in the synaptic compartment, a fraction far exceeding the 5% reported in total mouse hippocampus (*28*). The phosphoproteome is much more dynamic than the proteome (50% of proteins oscillating on at least one site vs. only 5% of oscillating proteins). Similar to our previous finding in the liver (*3*), mean fold-changes of the phosphoproteome are more than three-fold higher than in the proteome. At 7,000 accurately quantified phosphopeptides, the depth of our analysis allowed in depth bioinformatic analysis, retrieving many known but also highlighting novel temporally regulated processes at synapses. Overall, the phases of cycling phosphorylation fall into two main clusters at the boundaries of activity-rest transitions, implying a major role of synaptic phosphorylation in this process.

Numerous kinases are expressed in the brain, and some of them have been also localized to the synaptic compartment (*29*). However, our study clearly now classifies the synapse as a major kinase hub. Indeed, we detect more than 100 phosphorylated kinases (20% of the total kinome) and found that 50% cycle on at least one phosphoresidue. Combining protein interaction network data with our quantitative phosphoproteome revealed that these phosphorylation changes regulate the activity of at least a subset of them.

Together, our predictive and experimental data suggest phosphorylation-dependent temporal segregation of long-term potentiation (LTP) and long-term depression (LTD). Kinases with higher activity at the end of the rest period are involved in LTP, such as CAMK and PRKC, which function directly downstream of NMDARs receptor Ca2+ signaling (*8*). In contrast, ABL2 and DCLK1, two kinases modulating structural synaptic plasticity (*18, 19*) peak in activity at the transition to the rest period. Lack of ABL2 leads to elevated NMDAR synaptic currents (*19*), therefore ABL2 activation at the beginning of the resting phase may mediate synaptic downscaling. At the beginning of the active phase, GSK3b is phosphorylated on its inhibitory site S389. This blocks LTD and GABAR trafficking (*30*), further indicating a role of

phosphorylation in the temporal segregation of LTP and LTD. Figure 7 summarizes the complex phosphorylation pattern of many key components of these processes.

Together, our data point towards an association of synaptic potentiation with wakefulness (activity) while synaptic downscaling with rest. This supports the general synaptic homeostasis hypothesis for sleep – that synaptic downscaling is a critical function of sleep (*31*) at the molecular level and provides starting points for a multitude of mechanistic investigations.

To directly test the contribution of sleep and circadian cycles on phospho-dependent synaptic function, we sleep deprived mice at different times during the rest period. We observed that sleep pressure almost entirely abrogates global diurnal oscillation of phosphorylation in the synaptic compartment. This discovery establishes the importance of the homeostatic regulation of sleep over any other mechanism in generating daily rhythms of phosphorylation to modulate synaptic function.

Independent studies examining phosphorylation and sleep pressure support our findings. A very recent analysis in total mouse brain showed that increasing hours of sleep deprivation resulted in increasing average phosphorylation levels of 80 synaptic proteins (*10*). Based on these results, we examined our data and found phosphorylations in all of those 80 proteins. Our data also show rhythms over the day in 69 of them, although our site-specific data reveal a more nuanced picture than a simple overall increase in phosphorylation. Nevertheless, the fact that two completely different experimental approaches find concordant regulation on a subset of synaptic proteins provides validation, whereas our rhythmic data support the hypothesis that these phosphorylation sites are indeed part of the signatures of sleep.

Another recent phosphoproteomics study in the synaptic compartment showed that sleep regulates dephosphorylation of AMPA receptors in response to HOMER1a-dependent immediate early transcriptional signaling (*9*). Our data independently support these findings, extend them to several other neurotransmitter receptors and supply their diurnal rhythms of phosphorylation in response to sleep and activity pressure (**Fig. 7**).

Although the vast majority of protein phosphorylation appears to be regulated by sleep or wake states, a small number of circadian phosphorylations remain unchanged in amplitude and phase under sleep deprivation. These phosphorylations occur in a highly connected node of proteins involved in both synaptic vesicle trafficking (molecular motors, microtubule-associated proteins) and synaptic scaffold (SHANK3, PICCOLO, BASSOON), overall regulating excitatory synaptic strength, apparently irrespective of sleep-wake pressure. The fact that circadian rhythmicity regulates cortical function is already well documented at both behavioral and molecular levels (*32*). Together with our data, this makes it likely that even as sleep-wake pressures dynamically reconfigure synaptic structure and function, an underlying circadian rhythmicity continues at the molecular level, accounting for sleep-independent, circadian rhythms in cortical function.

In conclusion, this study represents the first *in vivo* evidence of orchestration of thousands of phosphorylation events in synapses across the day. These extensive phosphorylation rhythms temporally segregate synaptic activity in response to activity changes. Our study demonstrates a central role for rest-activity cycles in regulating phospho-dependent synaptic homeostasis in response to both sleep and wake pressure, potentially modulating inhibitory and excitatory synaptic plasticity. Interfering with rest-activity cycles almost completely abolished rhythms of phosphorylation in synaptic proteins, which may lead to dysfunctional synaptic plasticity associated with sleep deprivation.

References

1. D. Mauvoisin *et al.*, Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 167-172 (2014).
2. M. S. Robles, J. Cox, M. Mann, In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS genetics* **10**, e1004047 (2014).
3. M. S. Robles, S. J. Humphrey, M. Mann, Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology. *Cell Metab*, (2016).
4. F. Hosp, M. Mann, A Primer on Concepts and Applications of Proteomics in Neuroscience. *Neuron* **96**, 558-571 (2017).
5. K. Sharma *et al.*, Cell type- and brain region-resolved mouse brain proteome. *Nature neuroscience* **18**, 1819-1831 (2015).
6. J. J. Liu *et al.*, In vivo brain GPCR signaling elucidated by phosphoproteomics. *Science* **360**, (2018).
7. D. C. Dieterich, M. R. Kreutz, Proteomics of the Synapse--A Quantitative Approach to Neuronal Plasticity. *Molecular & cellular proteomics : MCP* **15**, 368-381 (2016).
8. K. M. Woolfrey, M. L. Dell'Acqua, Coordination of Protein Phosphorylation and Dephosphorylation in Synaptic Plasticity. *The Journal of biological chemistry* **290**, 28604-28612 (2015).
9. G. H. Diering *et al.*, Homer1a drives homeostatic scaling-down of excitatory synapses during sleep. *Science* **355**, 511-515 (2017).
10. Z. Wang *et al.*, Quantitative phosphoproteomic analysis of the molecular substrates of sleep need. *Nature*, (2018).
11. S. B. Noya *et al.*, The Cortical Synaptic Transcriptome is Organized by Clocks, but its Proteome is Driven by Sleep (2018).
12. P. R. Dunkley, P. E. Jarvie, P. J. Robinson, A rapid Percoll gradient procedure for preparation of synaptosomes. *Nature protocols* **3**, 1718-1728 (2008).
13. S. J. Humphrey, S. B. Azimifar, M. Mann, High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nature biotechnology* **33**, 990-995 (2015).
14. S. Tyanova *et al.*, The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13**, 731-740 (2016).
15. V. M. Ho, J. A. Lee, K. C. Martin, The cell biology of synaptic plasticity. *Science* **334**, 623-628 (2011).
16. K. O. Lai, Z. Liang, E. Fei, H. Huang, N. Y. Ip, Cyclin-dependent Kinase 5 (Cdk5)-dependent Phosphorylation of p70 Ribosomal S6 Kinase 1 (S6K) Is Required for Dendritic Spine Morphogenesis. *The Journal of biological chemistry* **290**, 14637-14646 (2015).
17. J. D. Rudolph, M. de Graauw, B. van de Water, T. Geiger, R. Sharan, Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. *Cell Syst* **3**, 585-593 e583 (2016).
18. E. Shin *et al.*, Doublecortin-like kinase enhances dendritic remodelling and negatively regulates synapse maturation. *Nature communications* **4**, 1440 (2013).
19. X. Xiao, A. D. Levy, B. J. Rosenberg, M. J. Higley, A. J. Koleske, Disruption of Coordinated Presynaptic and Postsynaptic Maturation Underlies the Defects in Hippocampal Synapse Stability and Plasticity in Abl2/Arg-Deficient Mice. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **36**, 6778-6791 (2016).
20. C. Cirelli, C. M. Gutierrez, G. Tononi, Extensive and divergent effects of sleep and wakefulness on brain gene expression. *Neuron* **41**, 35-43 (2004).

21. V. V. Vyazovskiy, C. Cirelli, M. Pfister-Genskow, U. Faraguna, G. Tononi, Molecular and electrophysiological evidence for net synaptic potentiation in wake and depression in sleep. *Nature neuroscience* **11**, 200-208 (2008).
22. I. Tobler, T. Deboer, M. Fischer, Sleep and sleep regulation in normal and prion protein-deficient mice. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **17**, 1869-1879 (1997).
23. S. Maret *et al.*, Homer1a is a core brain molecular correlate of sleep loss. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20090-20095 (2007).
24. N. Z. Gerges *et al.*, Independent functions of hsp90 in neurotransmitter release and in the continuous synaptic cycling of AMPA receptors. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **24**, 4758-4766 (2004).
25. D. Ivanova, A. Dirks, A. Fejtova, Bassoon and piccolo regulate ubiquitination and link presynaptic molecular dynamics with activity-regulated gene expression. *J Physiol* **594**, 5441-5448 (2016).
26. J. Li *et al.*, Long-term potentiation modulates synaptic phosphorylation networks and reshapes the structure of the postsynaptic interactome. *Science signaling* **9**, rs8 (2016).
27. G. Tononi, C. Cirelli, Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* **81**, 12-34 (2014).
28. C. K. Chiang *et al.*, Phosphoproteome Profiling Reveals Circadian Clock Regulation of Posttranslational Modifications in the Murine Hippocampus. *Front Neurol* **8**, 110 (2017).
29. L. L. Baltussen, F. Rosianu, S. K. Ultanir, Kinases in synaptic development and neurological diseases. *Prog Neuropsychopharmacol Biol Psychiatry* **84**, 343-352 (2018).
30. C. A. Bradley *et al.*, A pivotal role of GSK-3 in synaptic plasticity. *Front Mol Neurosci* **5**, 13 (2012).
31. G. Tononi, C. Cirelli, Sleep function and synaptic homeostasis. *Sleep Med Rev* **10**, 49-62 (2006).
32. R. Iyer, T. A. Wang, M. U. Gillette, Circadian gating of neuronal functionality: a basis for iterative metaplasticity. *Front Syst Neurosci* **8**, 164 (2014).
33. J. A. Vizcaino *et al.*, 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**, D447-456 (2016).
34. P. R. Dunkley, P. E. Jarvie, P. J. Robinson, A rapid Percoll gradient procedure for preparation of synaptosomes. *Nature protocols* **3**, 1718-1728 (2008).
35. M. S. Robles, S. J. Humphrey, M. Mann, Phosphorylation Is a Central Mechanism for Circadian Control of Metabolism and Physiology. *Cell Metab*, (2016).
36. S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols* **11**, 2301-2319 (2016).
37. S. Tyanova *et al.*, The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13**, 731-740 (2016).
38. M. S. Robles, J. Cox, M. Mann, In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS genetics* **10**, e1004047 (2014).

Acknowledgments: We thank K. Mayr, I. Paron and G. Sowa for technical assistance with the MS measurements, B. Collins for critical reading of the manuscript and B. Splettstößer for technical help with experimental workflow. **Funding:** This work was supported by the Max-Planck Society for the Advancement of Sciences and the German Research Foundation (DFG/Gottfried Wilhelm Leibniz Prize) funded this study. S.A.B. and S.B.N. were supported by the Swiss National Science Foundation, the Velux Foundation, the Human Frontiers Science Program, and the Zürich Clinical Research Priority Project “Sleep and Health” and are members

of the Neurosciences program within the Life Sciences Zürich Graduate School; **Author contributions:** S.B.N., M.S.R. and S.A.B. conceived and initiated the project and designed experiments; S.B.N., M.S.R. and F.B performed sample preparation and mass spectrometry experiments; M.S.R. and F.B. performed bioinformatic and data analysis with help from S.B.N. and T.B.; J.D.R. performed PHOTON analysis under supervision of J.C.; S.B.N., F.B., M.S.R., S.A.B. and M.M. wrote the manuscript with editing and input from T.B. and J.D.R. **Data and materials availability:** The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (*33*) partner repository with the dataset identifier PXD010697.

Supplementary Materials:

Materials and Methods

Figures S1-S5

Tables S1-S5

References # (34-38)

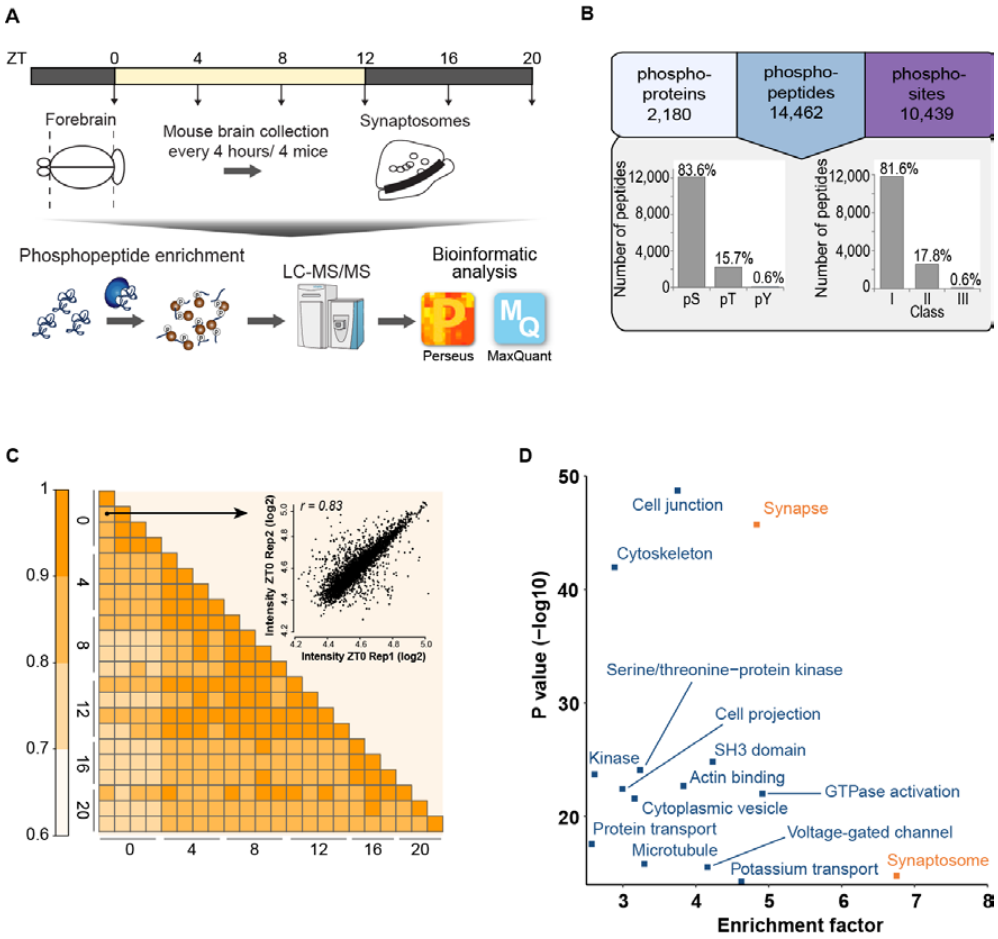


Fig. 1 Phosphoproteome characterization of synaptoneurosomes isolated across the day from mouse forebrains.

(A) Experimental workflow. **(B)** Number of identified phosphoproteins, phosphopeptides and phosphosites in all measured samples. Lower left: distribution of phosphorylated amino acids (Serine (pS), Threonine (pT) and Tyrosine (pY)). Lower right: number of phosphorylated residues from different classes according to localization probability -- Class I (probability > 75%), Class II (probability = 50 - 75%) and Class III (localization probability < 50%). **(C)** Heatmap representation of Pearson correlation coefficients calculated among phosphoproteomes. Inset: Example scatterplot of phosphopeptide intensities between two biological replicates of ZT0. **(D)** Scatterplot shows protein annotations (UniprotKB keywords) statistically enriched (Fisher's exact test FDR < 0.02) in the total synaptoneurosome phosphoproteome compared to an in-silico mouse gene list.

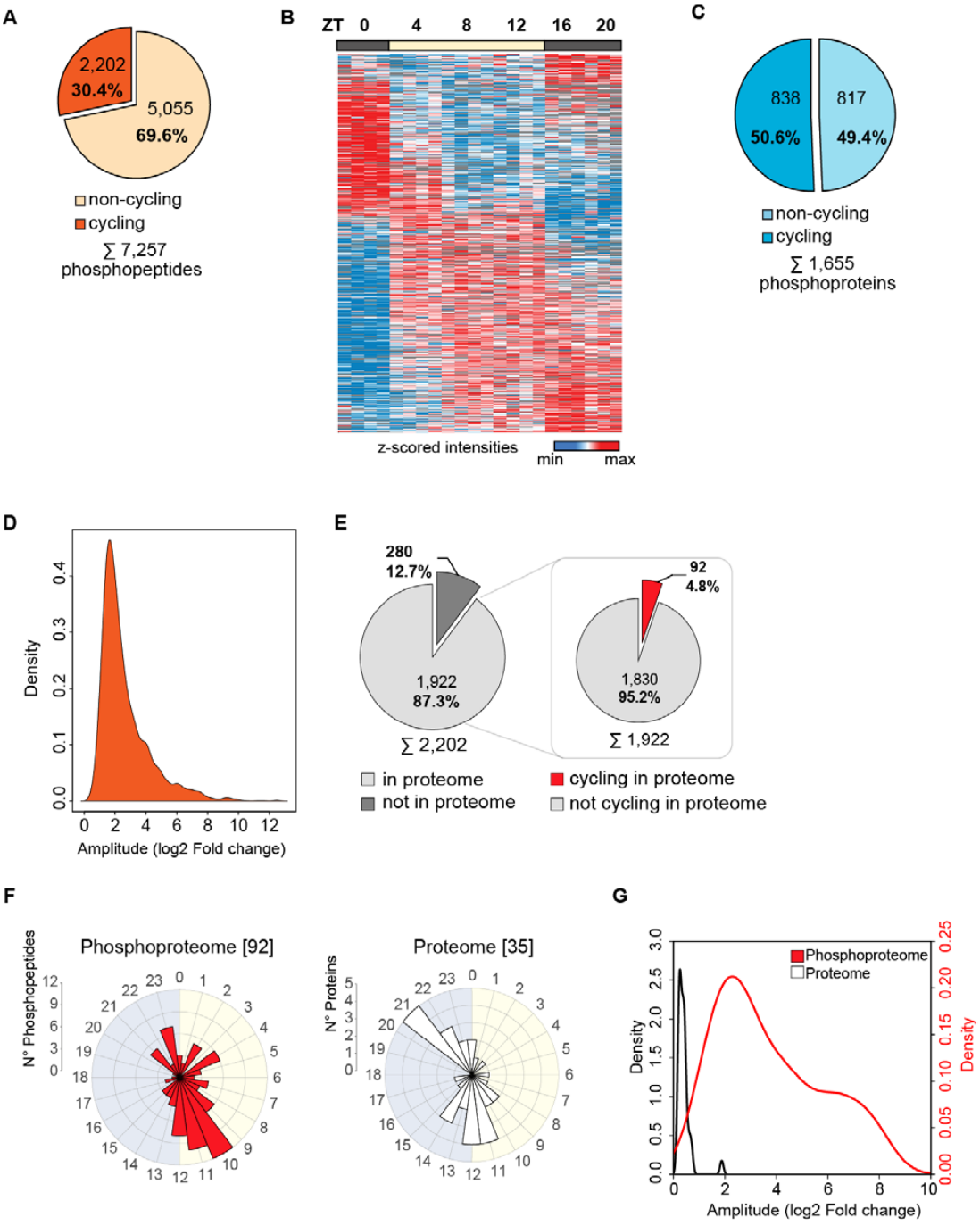

A
2,202
30.4%
5,055
69.6%
non-cycling
cycling
∑ 7,257
phosphopeptides
B
ZT 0 4 8 12 16 20
z-scored intensities
min max
C
838
50.6%
817
49.4%
non-cycling
cycling
∑ 1,655
phosphoproteins
D
Density
0.0 0.1 0.2 0.3 0.4
0 2 4 6 8 10 12
Amplitude (log2 Fold change)
E
280
12.7%
1,922
87.3%
∑ 2,202
92
4.8%
1,830
95.2%
∑ 1,922
in proteome
not in proteome
cycling in proteome
not cycling in proteome
F
Phosphoproteome [92]
N° Phosphopeptides
Proteome [35]
N° Proteins
G
Phosphoproteome
Proteome
Density
Density
Amplitude (log2 Fold change)

Fig. 2 Daily rhythms of the synaptic phosphoproteome.

(A) Pie chart showing the percentage of phosphopeptides oscillating daily in synaptoneurosomes. (**B)** Heat map with the intensities (log2 z-scored) of each cycling phosphopeptide (rows) across the measure samples (columns) ordered by peak of abundance. **(C)** Pie chart with the percentage of phosphoproteins oscillating daily in synaptoneurosomes. **(D)** Density plot showing the calculated amplitudes of rhythmic phosphopeptides in synaptoneurosomes. **(E)** Pie charts showing the percentage of cycling phosphopeptides from proteins quantified in our proteome study (left) and the fraction of cycling phosphopeptides from rhythmic proteins (right). **(F)** Rose plots representing the phase distribution of rhythmic phosphopeptides (left) and their corresponding oscillating proteins (right). **(G)** Density plots comparing the amplitudes of the rhythmic phosphopeptides and the corresponding oscillating proteins.

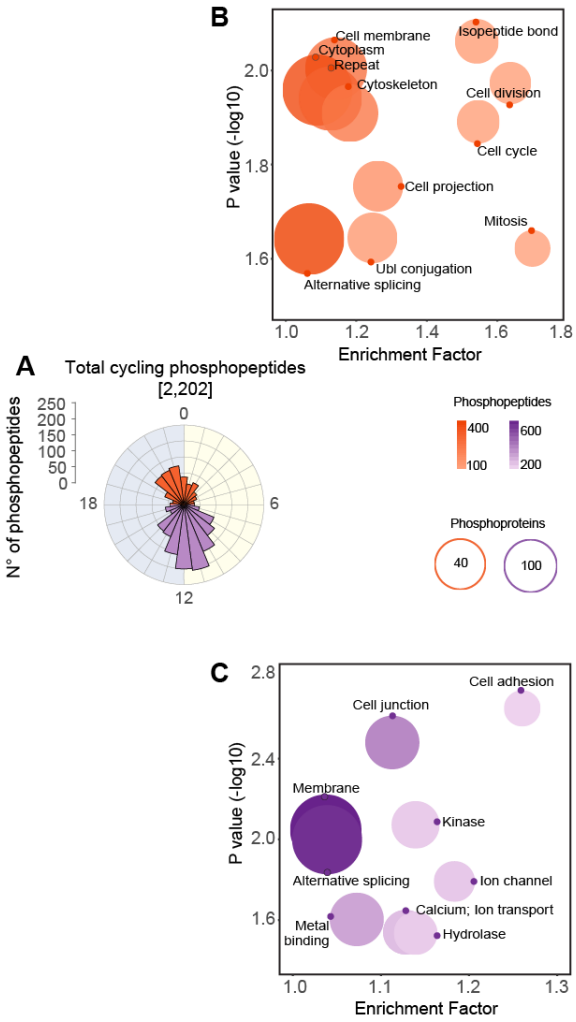


Fig. 3 Phosphorylation-dependent temporal control of synaptic functions.

(A) Rose plot showing the distribution of phases from cycling phosphopeptides of the total phosphoproteome. **(B)** Scatter plot showing the significantly enriched (Fisher's exact test $p < 0.05$) Uniprot Keywords protein annotations showing peak of phosphorylation cycles in the activity-rest transition (ZT18 to ZT6). Size of geometric points is proportional to the number of cycling phosphoproteins in the annotation and color intensity to the total number of phosphosites. **(C)** As in B but for the phosphopeptide cluster in the transition of rest-activity (ZT6 to ZT18).

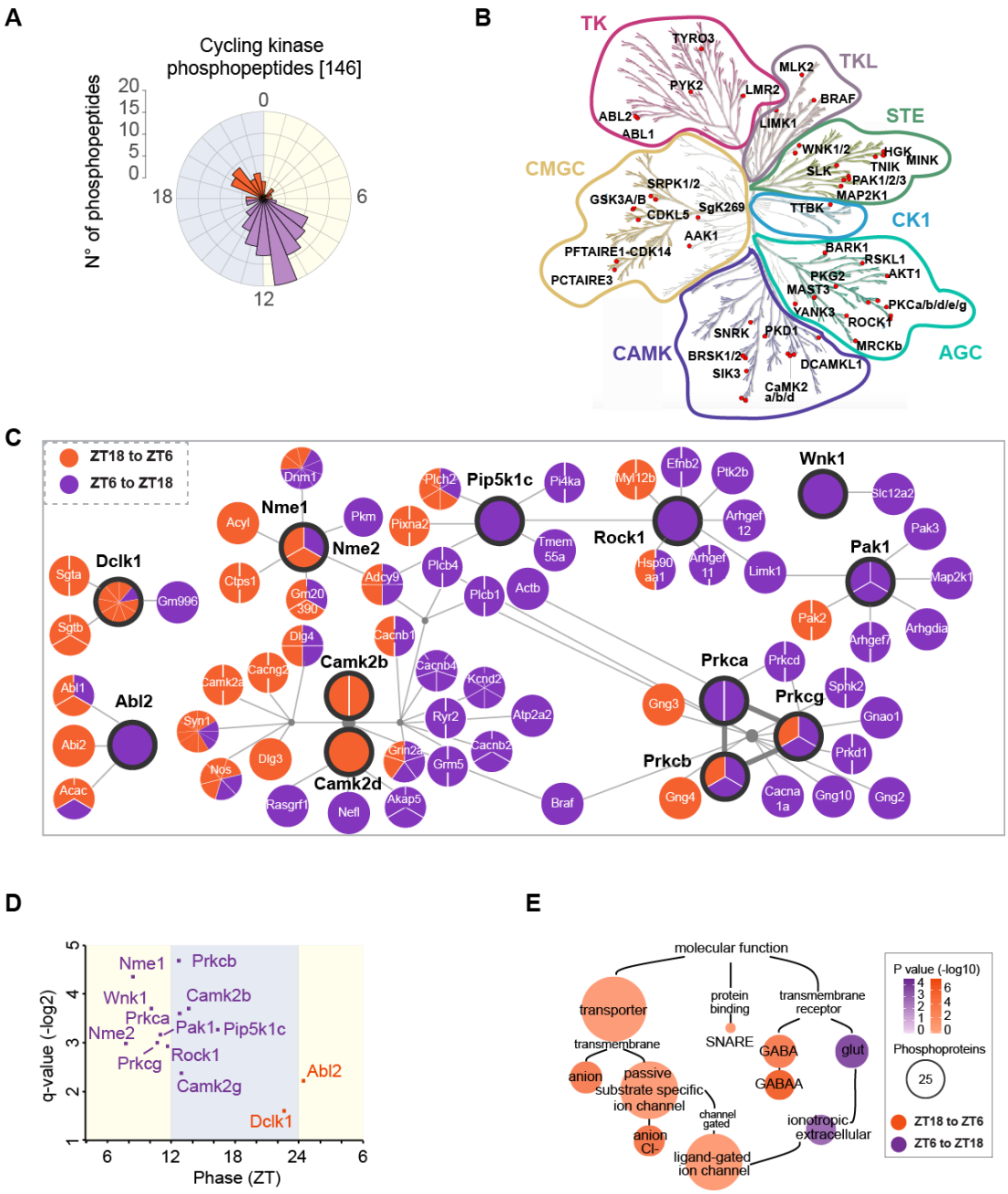

A
Cycling kinase phosphopeptides [146]
N° of phosphopeptides
20
15
10
5
0
0
6
12
18
B
TK
TYRO3
PYK2
ABL2
ABL1
TKL
MLK2
LMR2
BRAF
LIMK1
STE
WNK1/2
HGK
TNIK
MINK
SLK
PAK1/2/3
MAP2K1
TTBK
CMGC
SRPK1/2
GSK3A/B
CDKL5
SgK269
AAK1
PFTAIRE1-CDK14
PCTAIRE3
CK1
BARK1
RSKL1
AKT1
PKG2
MAST3
YANK3
PKCa/b/d/e/g
ROCK1
SNRK
PKD1
MRCKb
BRSK1/2
DCAMKL1
SIK3
CaMK2 a/b/d
CAMK
AGC
C
ZT18 to ZT6
ZT6 to ZT18
Dnm1
Nme1
Acyl
Pkm
Nme2
Pixna2
Plch2
Pip5k1c
Pi4ka
Tmem 55a
Myl12b
Efnb2
Ptk2b
Arhgef 12
Rock1
Hsp90 aa1
Arhgef 11
Limk1
Wnk1
Slc12a2
Pak3
Pak1
Map2k1
Arhgdia
Arhgef7
Pak2
Dclk1
Sgta
Sgtb
Gm996
Ctps1
Gm20 390
Adcy9
Plcb4
Plcb1
Actb
Dlg4
Cacnb1
Camk2b
Camk2a
Cacng2
Cacnb4
Kcnd2
Ryr2
Atp2a2
Cacnb2
Grin2a
Grm5
Camk2d
Akap5
Nefl
Rasgrf1
Braf
Abl2
Abl1
Abi2
Acac
Syn1
Nos
Dlg3
Prkca
Prkcd
Sphk2
Prkcg
Gnao1
Prkd1
Prkcb
Gng3
Gng4
Cacna 1a
Gng10
Gng2
D
q-value (-log2)
Phase (ZT)
Nme1
Prkcb
Wnk1
Camk2b
Prkca
Pak1
Pip5k1c
Nme2
Rock1
Prkcg
Camk2g
Abl2
Dclk1
E
molecular function
transporter
protein binding
transmembrane receptor
transmembrane
SNARE
GABA
glut
anion
passive
substrate specific
ion channel
GABAA
channel gated
ionotropic
extracellular
anion Cl-
ligand-gated ion channel
P value (-log10)
Phosphoproteins
25
ZT18 to ZT6
ZT6 to ZT18

Fig. 4 Rhythmic phosphorylation and activation of synaptic kinases.

(A) Rose plot with the phases of cycling phosphopeptides in kinases of synaptoneurosomes. Colors denote the clusters of activity-rest transition (ZT18-ZT6, orange) and rest-activity transition (ZT6-ZT18, purple). **(B)** Kinases with at least one phosphorylation cycling annotated to the major kinase families (52 out of the total 66 cycling) using http://www.kinhub.org. Tyrosine Kinases (TK), Tyrosine Kinase-Like (TKL), Homologs of the yeast STE7, STE11 and STE20 genes (STE), Casein/cell kinase 1 family (CK1), Protein Kinase A, G, C families (AGC), Calmodulin/Calcium regulated kinases and some non-calcium regulated families (CAMK), CDK, MAPK, GSK3 and CLK kinase families (CMGC). Atypical kinases are not shown. **(C)** Protein interaction network of PHOTON predicted cycling kinases ($q < 0.05$) with rhythmic phosphorylations with their cycling phosphorylated interactors. Nodes are divided based on the number of cycling phosphorylations color coded based on the clusters in A. **(D)** Scatter plot showing the PHOTON predicted peak of activation of kinases from C color coded based on the transition clusters shown in A. Phases are represented in the X-axis and the q-value obtained in the cycling analysis of the PHOTON scores in the Y-axis. **(E)** Representation of the GO Molecular Function annotations statistically enriched in a phase dependent manner using the PHOTON scores (phase dependent enrichment test $q < 0.05$, see Methods). Color indicates the phase of the annotation based on the cluster colors of A. Size of circles is proportional to the number of cycling phosphoproteins in the annotation and color intensity to the p-value of the enrichment test.

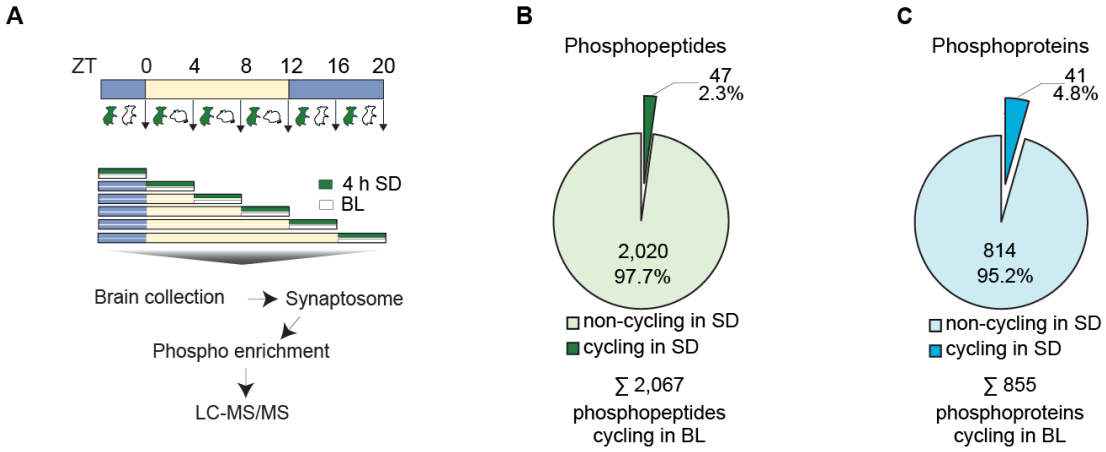


Fig. 5 Sleep deprivation abrogates synaptic phosphorylation rhythms.

(A) Experimental workflow used to profile the synaptoneurosome phosphoproteome under sleep deprivation across 24h. Sleep deprived (SD) animals were kept awake for 4h (green window) before being euthanized together with base line controls (BL). Blue and yellow indicate the light conditions during the protocol. Note that SD animals are awake regardless of the light conditions whereas BL mice rest during the light phase and are active in the dark. **(B)** Pie chart representing the fraction of cycling phosphopeptides (period=24h, $q<0.05$) in forebrain synaptoneurosomes from SD mice out of the 2,067 rhythmic in BL mice. **(C)** Pie chart representing the fraction of proteins with at least one cycling (period=24h, $q<0.05$) phosphopeptide in synaptoneurosomes of SD mice out of the 855 oscillating under BL conditions.

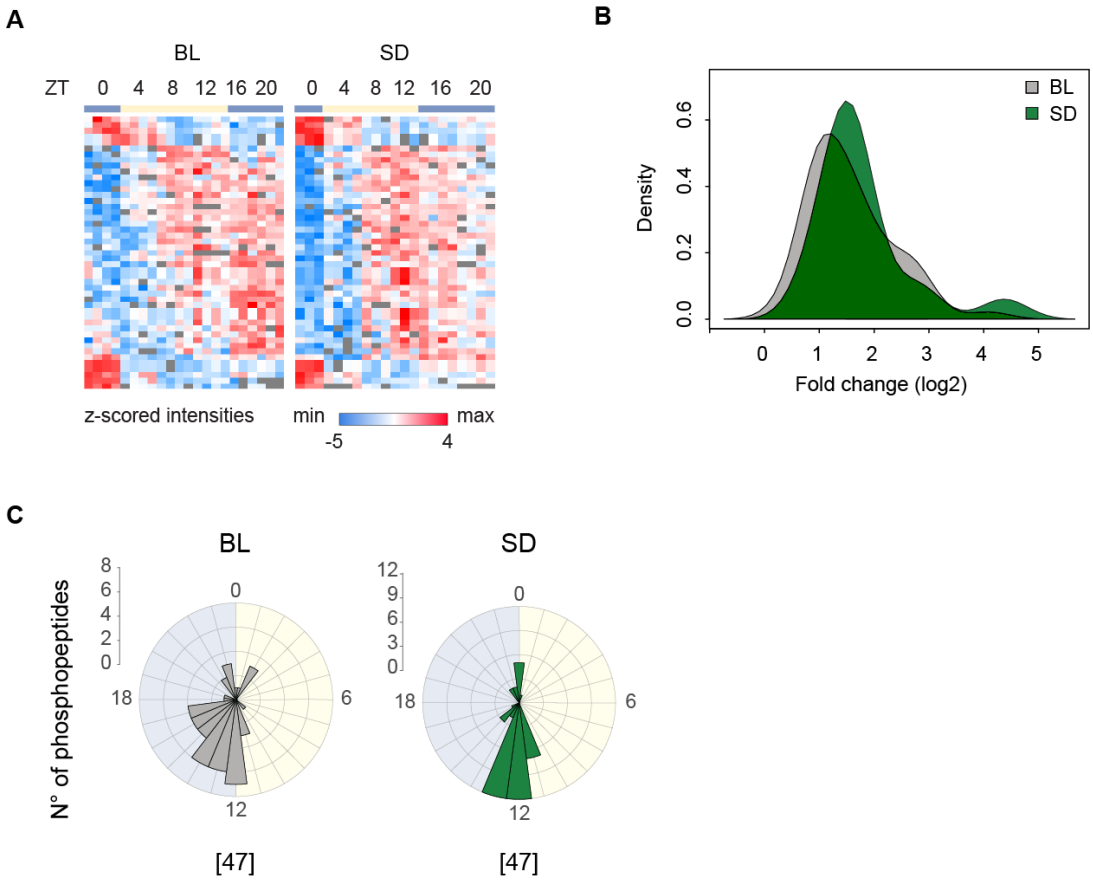


Fig. 6 Synaptic phosphoproteome cycling under sleep deprivation.

(A) Heat maps of the intensities (z-scored log2) of the 47 cycling phosphopeptides in BL (left) and SD (right) synaptoneurosomes ordered by peak of abundance. **(B)** Density plot with the fold change of the 47 phosphopeptides rhythmic in both conditions, calculated for BL (grey) and SD (green) conditions. **(C)** Rose plots showing the phase distribution in BL (left) and SD (right) for the 47 cycling phosphopeptides.

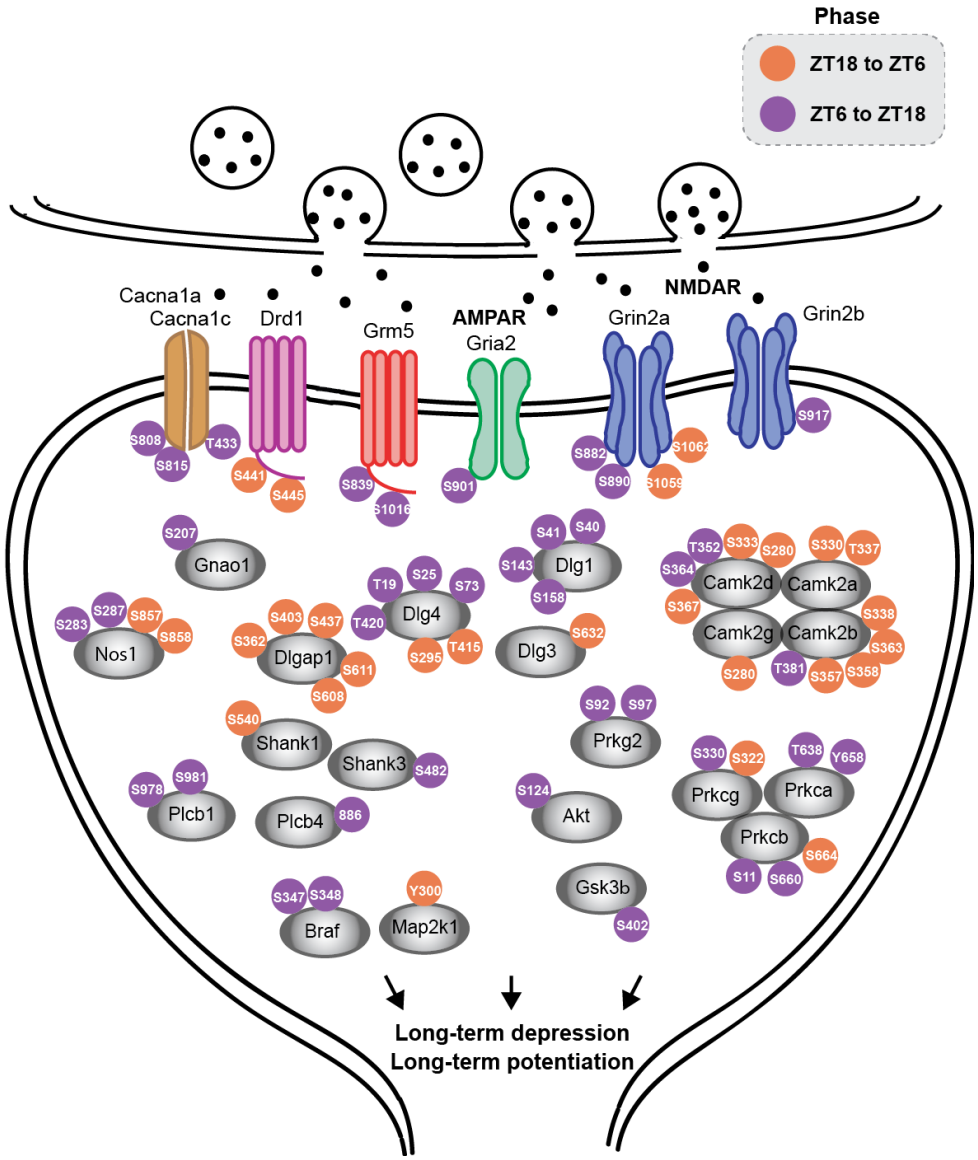


Fig. 7 Dynamic phosphorylation of synaptic plasticity mediators.

Schematic map of an excitatory synapse showing rhythmic phosphorylated sites detected in proteins involved in long term depression and potentiation. Number in the circle represents the phosphorylated amino acid and the color refers to the peak of the phosphorylation cycle as in Fig. 3A (orange from ZT18 to ZT6 and purple from ZT6 to ZT18).

2.3 MaxQuant goes Linux

MaxQuant [Cox and Mann, 2008] has been successfully used to analyze proteomics data for 10 years. Rapidly increasing dataset sizes have made the use of large servers running Linux, rather than workstations, more appealing. By adapting MaxQuant, which was initially written for the Windows-only .NET Framework, to the cross-platform Mono framework we enable more people to run MaxQuant on the computational resources available to them. I contributed to this joint lab effort by researching the required changes for the port of MaxQuant to Mono, and solving multiple user interface bugs, such as wrong interface scaling on high resolution monitors.

Pavel Sinitcyn, Shivani Tiwary, Jan Daniel Rudolph, Petra Gutenbrunner, Christoph Wichmann, Şule Yılmaz, Hamid Hamzeiy, Favio Salinas, and Jürgen Cox. MaxQuant goes Linux. *Nature methods*, 15:401, June 2018b. ISSN 1548-7105. URL `https://www.nature.com/articles/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.

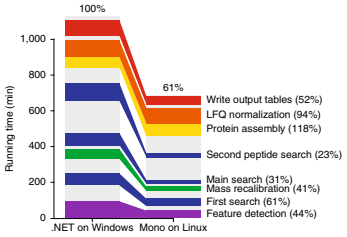


Fig. 1 | Benchmarking MaxQuant on Linux and Windows. We analyzed 300 LC-MS runs with MaxQuant using 120 logical cores in parallel, once with Ubuntu Linux (version 16.04.3) and once with Windows server 2012 R2 as the operating system. We used identical hardware in both cases: four Intel Xeon E7-4870 CPUs 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.

Chapter 3

Discussion and Outlook

In this thesis I presented the Perseus network module for the analysis of various proteomics data. It is aimed to enable researchers to not only generate novel PPI networks from pull down screens, but also analyze already available large-scale networks side-by-side with proteomics and phosphoproteomics data. Furthermore, previously hard-to-use tools have become available by breaking the programming language barrier with new interoperability infrastructure. Phosphoproteomic data can be transformed into protein-level signaling functionality scores and a predicted signaling pathway. Co-expression analysis uncovers the functional modules encoded in the data.

While MS-based proteomics has already proven to be competitive to the Y2H system for the study of PPIs, several technological advancements could improve upon the current state of the art. Pull-down screens rely on the endogenous tagging of a large number of proteins. CRISPR-Cas9 technology could accelerate the generation of tagged protein libraries [Lackner et al., 2015] especially in combination with smaller tags, such as the split-GFP epitope tag [Kamiyama et al., 2016]. Alternatively, proximity labeling has been successfully applied to identify proteins withing the same cellular compartment or organell [Rhee et al., 2013, Roux et al., 2013]. Utilizing chemical cross linking in combination with MS has already been used to get a more detailed view on the interactions within a protein complex by identify distance constraints [Herzog et al., 2012]. Rather than studying a single protein or complex, proteome-wide cross linking and protein-correlation profiling promise to interrogate all protein binding simultaneously [Liu et al., 2015, Kristensen and Foster, 2014]. However, the coverage of such studies is still very limited.

Even in the presence of complete interaction networks the main challenge of integrating such networks with other proteomics data, such as protein and phosphorylation site abundances, remains. Despite the establishment of a number of tools, including the ones presented in this thesis, for the analysis of such data, researchers still consider the classical analysis of protein lists separate from any network-based analyses. Demystifying the large PPI databases that to-date remain hidden behind web interfaces by allowing non-bioinformaticians to investigate and manipulate them in an intuitive and transparent manner will be a main driver towards their wider spread use. Even for researchers interested in only a single protein, studying not only which interaction partners it has, but also calculating simple network measures, such as the proteins degree compared to the degree distribution of the network can already provide functional insights on the biology of the protein. A hub protein will have a high degree, while a low degree implies peripheral location.

Many PPIs have been shown to be condition specific, such as dependent on phosphorylation [Tudor et al., 2015] or co-localization and highly dynamic. These characteristics are rarely represented in PPI networks and databases. For most applications, such as the calculation of signaling functionality scores from phosphoproteomic and PPI data presented earlier, the network remains static in all measured conditions. The large effort required to measure large-scale interactomes such as [Hein et al., 2013] prohibits the generation of PPI networks for each of the myriad imaginable conditions. A more practicable way would be to obtain a complete interactome of all potential PPIs and filter out interaction in a condition-specific manner. Tissue-level resolution can be obtained by retaining only interactions between proteins known to be expressed in the tissue [Bossi and Lehner, 2009]. In the future, such methods could be extended by utilizing the wealth of omics data and machine learning for the prediction of condition-specific PPI networks [Will and Helms, 2016].

Tools for studying phosphoproteomic data in the context of physical interactions usually score proteins from aggregated phosphorylation site information. Kinase activity is often modeled as a function of the phosphorylation changes observed on the substrates of the kinase [Hernandez-Armenta et al., 2017]. Such models have to rely on the scarce site-specific kinase-substrate interactions reported in public databases [Hornbeck et al., 2015], potentially supplemented by predictions [Horn et al., 2014]. For organisms other than human, the scarcity of the data affects analyses not only directly,

but also limits the training data available for prediction tools. Overlaps in the kinase-substrate assignments convolute the mechanistic interpretation of the data.

When kinase-substrate model assumptions are relaxed to accommodate higher quality undirected PPI networks the enzymatic interpretation of the scores is lost in return for higher coverage [Rudolph et al., 2016]. Common to both models is the reliance on first degree neighbors potentially losing the information encoded in the rest of the network. Phosphatases which de-phosphorylate their substrates in a condition specific manner are much less studied than kinases and not considered in any of the analyses described.

In order to understand signaling on a mechanistic level, ultimately, the function of phosphorylation patterns on proteins have to be understood. Master phosphorylation sites that regulated the activity or localization of a protein have been identified by low-throughput experiments. Additional phosphorylation sites appear to be ignored, similar to the 'junk DNA' hypothesis [Ohno, 1972]. While the 'junk DNA' hypothesis has been thoroughly debunked [Pennisi, 2012], large-scale characterization of phosphorylation sites has been limited to evolutionary conservation studies [Collins, 2009]. Going forward, linking kinase or signaling scores to the phosphorylation patterns observed on the protein itself, could provide an avenue for decoding the observed pattern.

In conclusion, this thesis provides a way forward for every researcher to leverage the power of the joint analysis of expression-omics tables and interaction networks.

Acronyms

BAC bacterial artificial chromosome

DDA data-dependent acquisision

DE differential expression

DIA data-independent acquisition

FDR false discovery rate

FTICR Fourier transform ion cyclotron

GBA guilt by association

GFP green fluorescent protein

HPLC high performance liquid chromatography

IMAC immobilized metal affinity chromatography

IP immunoprecipitation

LC liquid chromatography

MALDI matrix-assisted laser desorption/ionization

MS mass spectrometry

MS^1 full scan

MS^2 fragment scan

PCA principal component analysis

PDI protein-DNA interaction

PPI protein-protein interaction

PTM post-translational modification

RP reversed-phase

SCX strong cation exchange

SILAC stable isotope labeling by amino acids in cell culture

TAP tandem affinity purification

TMT tandem mass tag

TOF time-of-flight

TOM topological overlap measure

Y2H yeast two-hybrid

Bibliography

D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177–82, January 1987. ISSN 0036-8075. doi: 10.1126/SCIENCE.3798106. URL http://www.ncbi.nlm.nih.gov/pubmed/3798106.

G. W. Beadle and E. L. Tatum. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences*, 27(11):499–506, November 1941. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/16588492.

Albert László Barabási and Zoltán N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. ISSN 1471-0056. doi: 10.1038/nrg1272.

Ariel Bensimon, Albert J. R. Heck, and Ruedi Aebersold. Mass Spectrometry–Based Proteomics and Network Biology. *Annual Review of Biochemistry*, 81:379–405, 2012. ISSN 1545-4509. doi: 10.1146/annurev-biochem-072909-100424.

Claus Jørgensen and Marie Locard-Paulet. Analysing signalling networks by mass spectrometry. *Amino Acids*, 43(3):1061–1074, September 2012. ISSN 0939-4451. doi: 10.1007/s00726-012-1293-z.

Chunaram Choudhary and Matthias Mann. Decoding signalling networks by mass spectrometry-based proteomics. *Nature Reviews Molecular Cell Biology*, 11(6):427–439, June 2010. ISSN 1471-0072. doi: 10.1038/nrm2900. URL http://www.nature.com/articles/nrm2900.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002. ISSN 0034-6861. doi: 10.1103/RevModPhys.74.47.

B. Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–4, February 1998. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/9476889.

J. V. Olsen, M. Vermeulen, A. Santamaria, C. Kumar, M. L. Miller, L. J. Jensen, F. Gnad, J. Cox, T. S. Jensen, E. A. Nigg, S. Brunak, and M. Mann. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science Signaling*, 3(104):ra3, 2010. doi: 10.1126/scisignal.2000475.

Maria S. Robles, Jürgen Cox, and Matthias Mann. In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. *PLoS Genetics*, 10(1):e1004047, 2014. ISSN 1553-7390. doi: 10.1371/journal.pgen.1004047.

J. D. Jordan, E. M. Landau, and R. Iyengar. Signaling networks: the origins of cellular multitasking. *Cell*, 103(2):193–200, October 2000. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/11057893.

Marco Y. Hein, Nina C. Hubner, Ina Poser, Jürgen Cox, Nagarjuna Nagaraj, Yusuke Toyoda, Igor A. Gak, Ina Weisswange, Jörg Mansfeld, Frank Buchholz, Anthony A. Hyman, and Matthias Mann. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell*, 163(3):712–723, October 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.09.053. URL http://www.ncbi.nlm.nih.gov/pubmed/26496610.

Susan L. Kloet, Matthew M. Makowski, H. Irem Baymaz, Lisa van Voorthuijsen, Ino D. Karemaker, Alexandra Santanach, Pascal W. T. C. Jansen, Luciano Di Croce, and Michiel Vermeulen. The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nature Structural & Molecular Biology*, 23(7):682–690, July 2016. doi: 10.1038/nsmb.3248. URL http://www.ncbi.nlm.nih.gov/pubmed/27294783.

Marco Y. Hein, Kirti Sharma, Jürgen Cox, and Matthias Mann. Proteomic Analysis of Cellular Systems. *Handbook of Systems Biology*, pages 3–25, January 2013. doi: 10.1016/B978-0-12-385944-0.00001-0. URL https://www.sciencedirect.com/science/article/pii/B9780123859440000010?via{%}3Dihub.

Yaoyang Zhang, Bryan R. Fonslow, Bing Shan, Moon-Chang Baek, and John R. Yates III. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews*, 113(4):2343–2394, 2013.

A. Shevchenko, M. Wilm, O. Vorm, and M. Mann. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Analytical Chemistry*, 68(5):850–8, March 1996. ISSN 0003-2700. URL `http://www.ncbi.nlm.nih.gov/pubmed/8779443`.

Jacek R. Wiśniewski, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. Universal sample preparation method for proteome analysis. *Nature Methods*, 6(5):359–362, May 2009. ISSN 1548-7091. doi: 10.1038/nmeth.1322. URL `http://www.nature.com/articles/nmeth.1322`.

Nils A. Kulak, Garwin Pichler, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods*, 11(3):319–324, March 2014. ISSN 1548-7091. doi: 10.1038/nmeth.2834. URL `http://www.ncbi.nlm.nih.gov/pubmed/24487582`.

D. A. Wolters, M. P. Washburn, and J. R. Yates. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*, 73(23):5683–5690, 2001. ISSN 0003-2700. doi: 10.1021/ac010617e.

James W. Jorgenson. Capillary Liquid Chromatography at Ultrahigh Pressures. *Annual Review of Analytical Chemistry*, 3(1):129–150, June 2010. ISSN 1936-1327. doi: 10.1146/annurev.anchem.1.031207.113014.

J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.2675315.

Michael. Karas and Franz. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60(20):2299–2301, October 1988. ISSN 0003-2700. doi: 10.1021/ac00171a028.

Michaela Scigelova, Martin Hornshaw, Anastassios Giannakopulos, and Alexander Makarov. Fourier transform mass spectrometry. *Molecular & Cellular Proteomics*, 10

(7):M111.009431, July 2011. ISSN 1535-9484. doi: 10.1074/mcp.M111.009431. URL http://www.ncbi.nlm.nih.gov/pubmed/21742802.

Roman A Zubarev and Alexander Makarov. Orbitrap mass spectrometry. *Analytical chemistry*, 85:5288–5296, June 2013. ISSN 1520-6882. doi: 10.1021/ac4001223.

Richard Alexander Scheltema, Jan-Peter Hauschild, Oliver Lange, Daniel Hornburg, Eduard Denisov, Eugen Damoc, Andreas Kuehn, Alexander Makarov, and Matthias Mann. The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Molecular & Cellular Proteomics*, 13(12):3698–708, December 2014. ISSN 1535-9484. doi: 10.1074/mcp.M114.043489. URL http://www.ncbi.nlm.nih.gov/pubmed/25360005.

Vivien Marx. Targeted proteomics. *Nature Methods*, 10(1):19–22, January 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2285. URL http://www.nature.com/articles/nmeth.2285.

Matthias Mann, Ronald C. Hendrickson, and Akhilesh Pandey. Analysis of Proteins and Proteomes by Mass Spectrometry. *Annual Review of Biochemistry*, 70(1):437–473, June 2001. ISSN 0066-4154. doi: 10.1146/annurev.biochem.70.1.437.

Ludovic C. Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*, 11(6):O111.016717, June 2012. ISSN 1535-9484. doi: 10.1074/mcp.O111.016717. URL http://www.ncbi.nlm.nih.gov/pubmed/22261725.

Jürgen Cox, Marco Y. Hein, Christion A. Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics*, 13(9):2513–2526, 2014. ISSN 1535-9484. doi: 10.1074/mcp.M113.031591.

Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.

Molecular & Cellular Proteomics, 1(5):376–86, May 2002. ISSN 1535-9476. doi: 10.1074/MCP.M200025-MCP200. URL http://www.ncbi.nlm.nih.gov/pubmed/12118079.

Blagoy Blagoev, Shao-En Ong, Irina Kratchmarova, and Matthias Mann. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature Biotechnology*, 22(9):1139–1145, September 2004. ISSN 1087-0156. doi: 10.1038/nbt1005. URL http://www.nature.com/articles/nbt1005.

Andrew Thompson, Jürgen Jurgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Gunter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: a novel quantificaiton strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003. ISSN 0003-2700. doi: 10.1021/ac0262560.

Thilo Werner, Gavain Sweetman, Maria Fälth Savitski, Toby Mathieson, Marcus Bantscheff, and Mikhail M. Savitski. Ion Coalescence of Neutron Encoded TMT 10-Plex Reporter Ions. *Analytical Chemistry*, 86(7):3594–3601, April 2014. ISSN 0003-2700. doi: 10.1021/ac500140s.

Mikhail M. Savitski, Toby Mathieson, Nico Zinn, Gavain Sweetman, Carola Doce, Isabelle Becher, Fiona Pachl, Bernhard Kuster, and Marcus Bantscheff. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of Proteome Research*, 12(8):3586–3598, 2013. ISSN 1535-3893. doi: 10.1021/pr400098r.

Jeremy D. O'Connell, Joao A. Paulo, Jonathon J. O'Brien, and Steven P. Gygi. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *Journal of Proteome Research*, 17(5):1934–1942, May 2018. ISSN 1535-3893. doi: 10.1021/acs.jproteome.8b00016. URL http://www.ncbi.nlm.nih.gov/pubmed/29635916.

Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, 2008. ISSN 1087-0156. doi: 10.1038/nbt.1511. URL http://www.ncbi.nlm.nih.gov/pubmed/19029910.

Stefka Tyanova, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y. Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9):731–740, June 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3901.

Pavel Sinitcyn, Jan Daniel Rudolph, and Jürgen Cox. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science*, 1(1):annurev–biodatasci–080917–013516, July 2018a. ISSN 2574-3414. doi: 10.1146/annurev-biodatasci-080917-013516.

G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(10):1030–1032, October 1999. ISSN 1087-0156. doi: 10.1038/13732. URL http://www.nature.com/articles/nbt1099{_}1030.

Oscar Puig, Friederike Caspary, Guillaume Rigaut, Berthold Rutz, Emmanuelle Bouveret, Elisabeth Bragado-Nilsson, Matthias Wilm, and Bertrand Séraphin. The tandem affinity purification (TAP) method: A general procedure of protein complex purification. *Methods*, 24(3):218–229, July 2001. ISSN 1046-2023. doi: 10.1006/meth.2001.1183. URL https://www.sciencedirect.com/science/article/pii/S1046202301911831?via{%}3Dihub.

Gareth Butland, José Manuel Peregrin-Alvarez, Joyce Li, Wehong Yang, Xiaochun Yang, Veronica Canadien, Andrei Starostine, Dawn Richards, Bryan Beattie, Nevan Krogan, Michael Davey, John Parkinson, Jack Greenblatt, and Andrew Emili. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, 433(7025):531–537, February 2005. ISSN 0028-0836. doi: 10.1038/nature03239. URL http://www.ncbi.nlm.nih.gov/pubmed/15690043.

Anne Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dümpelfeld, Angela Edelmann, Marie Anne Heurtier, Verena Hoffman, Christian Hoefert, Karin Klein, Manuela Hudak, Anne Marie Michon, Malgorzata Schelder, Markus Schirle, Marita Remor, Tatjana Rudi, Sean Hooper, Andreas Bauer, Tewis Bouwmeester, Georg Casari, Gerard Drewes, Gitte Neubauer, Jens M. Rick, Bernhard Kuster, Peer Bork, Robert B. Russell, and Giulio Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, March 2006. ISSN 0028-0836. doi: 10.1038/nature04532. URL http://www.nature.com/articles/nature04532.

Peter Uetz, Loic Glot, Gerard Cagney, Traci A. Mansfield, Richard S. Judson,

James R. Knight, Daniel Lockshon, Vaibhav Narayan, Malthreyan Srinivasan, Pascale Pochart, Alla Qureshi-Emlli, Ying Li, Brian Godwin, Diana Conover, Theodore Kalbfleisch, Govindan Vijayadamodar, Meijia Yang, Mark Johnston, Stanley Fields, and Jonathan M. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, February 2000. ISSN 0028-0836. doi: 10.1038/35001009. URL `http://www.ncbi.nlm.nih.gov/pubmed/10688190`.

Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E. Wanker. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6): 957–968, September 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.08.029. URL `https://www.sciencedirect.com/science/article/pii/S0092867405008664`.

Waltraud X. Schulze and Matthias Mann. A Novel Proteomic Screen for Peptide-Protein Interactions. *Journal of Biological Chemistry*, 279(11):10756–10764, March 2004. ISSN 0021-9258. doi: 10.1074/jbc.M309909200.

Eva C. Keilhauer, Marco Y. Hein, and Matthias Mann. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Molecular & Cellular Proteomics*, 14(1):120–35, January 2015. ISSN 1535-9484. doi: 10.1074/mcp.M114.041012. URL `http://www.ncbi.nlm.nih.gov/pubmed/25363814`.

Nina C. Hubner, Alexander W. Bird, Jürgen Cox, Bianca Splettstoesser, Peter Bandilla, Ina Poser, Anthony Hyman, and Matthias Mann. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *Journal of Cell Biology*, 189(4):739–754, May 2010. ISSN 0021-9525. doi: 10.1083/jcb.200911091. URL `http://www.ncbi.nlm.nih.gov/pubmed/20479470`.

Edward L. Huttlin, Lily Ting, Raphael J. Bruckner, Fana Gebreab, Melanie P. Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, Rui Dong, Virginia Guarani, Laura Pontano Vaites, Alban Ordureau, Ramin Rad, Brian K. Erickson, Mar-

tin Wühr, Joel Chick, Bo Zhai, Deepak Kolippakkam, Julian Mintseris, Robert A. Obar, Tim Harris, Spyros Artavanis-Tsakonas, Mathew E. Sowa, Pietro De Camilli, Joao A. Paulo, J. Wade Harper, and Steven P. Gygi. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2):425–440, July 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.06.043. URL `http://www.ncbi.nlm.nih.gov/pubmed/26186194`.

Edward L. Huttlin, Raphael J. Bruckner, Joao A. Paulo, Joe R. Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P. Gygi, Hannah Parzen, John Szpyt, Stanley Tam, Gabriela Zarraga, Laura Pontano-Vaites, Sharan Swarup, Anne E. White, Devin K. Schweppe, Ramin Rad, Brian K. Erickson, Robert A. Obar, K. G. Guruharsha, Kejie Li, Spyros Artavanis-Tsakonas, Steven P. Gygi, and J. Wade Harper. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, May 2017. ISSN 1476-4687. doi: 10.1038/nature22366. URL `http://www.ncbi.nlm.nih.gov/pubmed/28514442`.

Alexey I. Nesvizhskii, Olga Vitek, and Ruedi Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*, 4(10):787–797, October 2007. ISSN 1548-7091. doi: 10.1038/nmeth1088. URL `http://www.ncbi.nlm.nih.gov/pubmed/17901868`.

V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001. ISSN 0027-8424. doi: 10.1073/pnas.091062498.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57:289–300, 1995.

Hyungwon Choi, Brett Larsen, Zhen Yuan Lin, Ashton Breitkreutz, Dattatreya Mellacheruvu, Damian Fermin, Zhaohui S. Qin, Mike Tyers, Anne Claude Gingras, and Alexey I. Nesvizhskii. SAINT: Probabilistic scoring of affinity purificationg-mass spectrometry data. *Nature Methods*, 8(1):70–73, January 2011. ISSN 1548-7091. doi: 10.1038/nmeth.1541. URL `http://www.ncbi.nlm.nih.gov/pubmed/21131968`.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos,

Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian Von Mering. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, January 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1003. URL `http://www.ncbi.nlm.nih.gov/pubmed/25352553`.

Andrew Chatr-Aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379, 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw1102.

Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C. Lovering, Birgit Meldal, Anna N. Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim Van Roey, Gianni Cesareni, and Henning Hermjakob. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1), 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1115.

T. S. Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K. Kashyap, Riaz Mohmood, Y. L. Ramachandra, V. Krishna, B. Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(18988627):D767–D772, January 2009. ISSN 0305-1048. URL `https://www.ncbi.nlm.nih.gov/pmc/PMC2686490/`.

Nir Yosef, Lior Ungar, Einat Zalckvar, Adi Kimchi, Martin Kupiec, Eytan Ruppin, and Roded Sharan. Toward accurate reconstruction of functional protein networks.

Molecular Systems Biology, 5:248, 2009. ISSN 1744-4292. doi: 10.1038/msb.2009.3. URL http://www.ncbi.nlm.nih.gov/pubmed/19293828.

Gregorio Alanis-Lobato, Miguel A. Andrade-Navarro, and Martin H. Schaefer. HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414, 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw985.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1070.

Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, Lisa Matthews, Bruce May, Marija Milacic, Karen Rothfels, Veronica Shamovsky, Marissa Webber, Joel Weiser, Mark Williams, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1351.

Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H. Werner Mewes. CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research*, 38(SUPPL.1), October 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp914.

David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, June 2007. ISSN 0036-8075. doi: 10.1126/science.1141319. URL http://www.ncbi.nlm.nih.gov/pubmed/17540862.

Tijana Milenković and Natasa Przulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–73, 2008. ISSN 1176-9351. URL http://www.ncbi.nlm.nih.gov/pubmed/19259413.

Noël Malod-Dognin and Nataša Pržulj. L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics*, 31(13):2182–2189, July 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btv130.

Melissa S. Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, Kristina Hanspers, Ruth Isserlin, Ryan Kelley, Sarah Killcoyne, Samad Lotia, Steven Maere, John Morris, Keiichiro Ono, Vuk Pavlovic, Alexander R. Pico, Aditya Vailaya, Peng-Liang Wang, Annette Adler, Bruce R. Conklin, Leroy Hood, Martin Kuiper, Chris Sander, Ilya Schmulevich, Benno Schwikowski, Guy J. Warner, Trey Ideker, and Gary D. Bader. Integration of biological networks and gene expression data using Cytoscape. *Nature protocols*, 2(10):2366–82, 2007. ISSN 1750-2799. doi: 10.1038/nprot.2007.324. URL `http://www.ncbi.nlm.nih.gov/pubmed/17947979`.

Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–40, 2002. ISSN 1367-4803. URL `http://www.ncbi.nlm.nih.gov/pubmed/12169552`.

Nir Yosef, Einat Zalckvar, Assaf D. Rubinstein, Max Homilius, Nir Atias, Liram Vardi, Igor Berman, Hadas Zur, Adi Kimchi, Eytan Ruppin, and Roded Sharan. ANAT: A tool for constructing and analyzing functional protein networks. *Science Signaling*, 4 (196):pl1–pl1, 2011. ISSN 1945-0877. doi: 10.1126/scisignal.2001935.

Nurcan Tuncbag, Sara J. C. Gosline, Amanda Kedaigle, Anthony R. Soltis, Anthony Gitter, and Ernest Fraenkel. Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLOS Computational Biology*, 12(4):e1004879, April 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004879.

R Core Team. R: A Language and Environment for Statistical Computing, 2018. URL `https://www.r-project.org`.

Guido van Rossum. Python tutorial. Technical report, Amsterdam, 1995.

Max Franz, Christian T. Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D. Bader. Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, 32(2):309–311, September 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btv557.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker. Cytoscape: A software

Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303.

Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16):3448–3449, 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti551.

Quiagen Inc. IPA. URL `https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis`.

H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001. ISSN 0028-0836. doi: 10.1038/35075138. URL `http://www.nature.com/articles/35075138`.

P. Cohen. The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends in biochemical sciences*, 25(12):596–601, December 2000. ISSN 0968-0004. URL `http://www.ncbi.nlm.nih.gov/pubmed/11116185`.

T. Hunter. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell*, 80(2):225–36, January 1995. ISSN 0092-8674. URL `http://www.ncbi.nlm.nih.gov/pubmed/7834742`.

Boris Macek, Matthias Mann, and Jesper V. Olsen. Global and Site-Specific Quantitative Phosphoproteomics: Principles and Applications. *Annual Review of Pharmacology and Toxicology*, 49(1):199–221, February 2009. ISSN 0362-1642. doi: 10.1146/annurev.pharmtox.011008.145606.

Juan-José Ventura and Angel R. Nebreda. Protein kinases and phosphatases as therapeutic targets in cancer. *Clinical & translational oncology*, 8(3):153–60, March 2006. ISSN 1699-048X. URL `http://www.ncbi.nlm.nih.gov/pubmed/16648114`.

Nicholas M. Riley and Joshua J. Coon. Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. *Analytical Chemistry*, 88(1):74–94, 2016. ISSN 1520-6882. doi: 10.1021/acs.analchem.5b04123.

Jacek R. Wiśniewski and Matthias Mann. Consecutive Proteolytic Digestion in an Enzyme Reactor Increases Depth of Proteomic and Phosphoproteomic Analysis. *Analytical Chemistry*, 84(6):2631–2637, March 2012. ISSN 0003-2700. doi: 10.1021/ac300006b.

Mads Grønborg, Troels Zakarias Kristiansen, Allan Stensballe, Jens S. Andersen, Osamu Ohara, Matthias Mann, Ole Nørregaard Jensen, and Akhilesh Pandey. A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate. *Molecular & Cellular Proteomics*, 1(7):517–27, July 2002. ISSN 1535-9476. doi: 10.1074/MCP.M200010-MCP200. URL `http://www.ncbi.nlm.nih.gov/pubmed/12239280`.

Tine E. Thingholm, Thomas J. D. Jørgensen, Ole N. Jensen, and Martin R. Larsen. Highly selective enrichment of phosphorylated peptides using titanium dioxide. *Nature Protocols*, 1(4):1929–1935, November 2006. ISSN 1754-2189. doi: 10.1038/nprot.2006.185.

Judit Villén and Steven P. Gygi. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nature Protocols*, 3(10):1630–1638, October 2008. ISSN 1754-2189. doi: 10.1038/nprot.2008.150. URL `http://www.ncbi.nlm.nih.gov/pubmed/18833199`.

Sean J. Humphrey, S. Babak Azimifar, and Matthias Mann. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nature biotechnology*, 33(9):990–5, August 2015. ISSN 1546-1696. doi: 10.1038/nbt.3327.

Sean J. Humphrey, Ozge Karayel, David E. James, and Matthias Mann. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. *Nature Protocols*, 13(9):1897–1916, September 2018. ISSN 1754-2189. doi: 10.1038/s41596-018-0014-9. URL `http://www.ncbi.nlm.nih.gov/pubmed/30190555`.

Alexander Hogrebe, Louise von Stechow, Dorte B. Bekker-Jensen, Brian T. Weinert, Christian D. Kelstrup, and Jesper V. Olsen. Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nature Communications*, 9(1):1045, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03309-6. URL `http://www.nature.com/articles/s41467-018-03309-6`.

Clement M Potel, Simone Lemeer, and Albert J R Heck. Phosphopeptide fragmentation and site localization by mass spectrometry; an update. *Analytical chemistry*, November 2018. ISSN 1520-6882. doi: 10.1021/acs.analchem.8b04746.

S. A. Beausoleil, J. Villen, S. A. Gerber, J. Rush, and S. P. Gygi. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology*, 24(10):1285–1292, 2006. doi: 10.1038/nbt1240.

Jürgen Cox, Annette Michalski, and Matthias Mann. Software lock mass by two-dimensional minimization of peptide mass errors. *Journal of the American Society for Mass Spectrometry*, 22(8):1373–1380, 2011. ISSN 1044-0305. doi: 10.1007/s13361-011-0142-8.

Kirti Sharma, Rochelle C. J. D'Souza, Stefka Tyanova, Christoph Schaab, Jacek R. Wiśniewski, Jürgen Cox, and Matthias Mann. Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Reports*, 8(5):1583–1594, 2014. ISSN 2211-1247. doi: 10.1016/j.celrep.2014.07.036.

M. O. Collins. Evolving Cell Signals. *Science*, 325(5948):1635–1636, September 2009. ISSN 0036-8075. doi: 10.1126/science.1180331.

Jan Daniel Rudolph and Jürgen Cox. A network module for the Perseus software for computational proteomics facilitates proteome interaction graph analysis. *bioRxiv*, page 447268, October 2018. doi: 10.1101/447268. URL `https://www.biorxiv.org/content/early/2018/10/18/447268?rss=1`.

Peter V. Hornbeck, Bin Zhang, Beth Murray, Jon M. Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research*, 43(D1):D512–D520, January 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1267. URL `http://www.ncbi.nlm.nih.gov/pubmed/25514926`.

Daniel Schwartz and Steven P. Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology*, 23(11):1391–1398, November 2005. ISSN 1087-0156. doi: 10.1038/nbt1146. URL `http://www.ncbi.nlm.nih.gov/pubmed/16273072`.

Martin Lee Miller, Lars Juhl Jensen, Francesca Diella, Claus Jørgensen, Michele Tinti, Lei Li, Marilyn Hsiung, Sirlester A. Parker, Jennifer Bordeaux, Thomas Sicheritz-Ponten, Marina Olhovsky, Adrian Pasculescu, Jes Alexander, Stefan Knapp, Nikolaj Blom, Peer Bork, Shawn Li, Gianni Cesareni, Tony Pawson, Benjamin E. Turk, Michael B. Yaffe, Søren Brunak, and Rune Linding. Linear motif atlas for

phosphorylation-dependent signaling. *Science Signaling*, 1(35):ra2–ra2, September 2008. ISSN 1945-0877. doi: 10.1126/scisignal.1159433.

Rune Linding, Lars Juhl Jensen, Adrian Pasculescu, Marina Olhovsky, Karen Colwill, Peer Bork, Michael B. Yaffe, and Tony Pawson. NetworKIN: A resource for exploring cellular phosphorylation networks. *Nucleic Acids Research*, 36(SUPPL. 1), 2008. ISSN 0305-1048. doi: 10.1093/nar/gkm902.

Heiko Horn, Erwin M. Schoof, Jinho Kim, Xavier Robin, Martin L. Miller, Francesca Diella, Anita Palma, Gianni Cesareni, Lars Juhl Jensen, and Rune Linding. KinomeXplorer: An integrated platform for kinome biology studies. *Nature Methods*, 11(6):603–604, June 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2968. URL `http://www.nature.com/articles/nmeth.2968`.

Pedro Casado, Juan Carlos Rodriguez-Prados, Sabina C. Cosulich, Sylvie Guichard, Bart Vanhaesebroeck, Simon Joel, and Pedro R. Cutillas. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Science Signaling*, 6(268):rs6–rs6, 2013. ISSN 1945-0877. doi: 10.1126/scisignal.2003573.

Claudia Hernandez-Armenta, David Ochoa, Emanuel Gonçalves, Julio Saez-Rodriguez, and Pedro Beltrao. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, 33(12):1845–1851, 2017. ISSN 1460-2059. doi: 10.1093/bioinformatics/btx082.

Edmund H. Wilkes, Pedro Casado, Vinothini Rajeeve, and Pedro R. Cutillas. Kinase activity ranking using phosphoproteomics data (KARP) quantifies the contribution of protein kinases to the regulation of cell viability. *Molecular & Cellular Proteomics*, 16(9):1694–1704, 2017. ISSN 1535-9476. doi: 10.1074/mcp.O116.064360.

Marcel Mischnik, Francesca Sacco, Jürgen Cox, Hans Christoph Schneider, Matthias Schäfer, Manfred Hendlich, Daniel Crowther, Matthias Mann, and Thomas Klabunde. IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics*, 32(3):424–431, 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btv699.

Camille D. A. Terfve, Edmund H. Wilkes, Pedro Casado, Pedro R. Cutillas, and Julio Saez-Rodriguez. Large-scale models of signal propagation in human cells derived

from discovery phosphoproteomic data. *Nature Communications*, 6:8033, 2015. ISSN 2041-1723. doi: 10.1038/ncomms9033.

Jan Daniel Rudolph, Marjo de Graauw, Bob van de Water, Tamar Geiger, and Roded Sharan. Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. *Cell Systems*, 3(6):585–593.e3, December 2016. ISSN 2405-4720. doi: 10.1016/j.cels.2016.11.005. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471216303696.

Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, December 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-559. URL http://www.ncbi.nlm.nih.gov/pubmed/19114008.

Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Favera, and Andrea Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-S1-S7.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69:066138, June 2004. ISSN 1539-3755. doi: 10.1103/PhysRevE.69.066138.

Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1): 328, December 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-328.

Bin Zhang and Steve Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article17, 2005. ISSN 1544-6115. doi: 10.2202/1544-6115.1128. URL https://www.ncbi.nlm.nih.gov/pubmed/16646834.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, August 2002. ISSN 1095-9203. doi: 10.1126/science.1073374. URL http://www.ncbi.nlm.nih.gov/pubmed/12202830.

Stephen Oliver. Guilt-by-association goes global. *Nature*, 403(6770):601–602, February

2000. ISSN 0028-0836. doi: 10.1038/35001165. URL http://www.ncbi.nlm.nih.gov/pubmed/10688178.

H. K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Research*, 14(6):1085–1094, May 2004. ISSN 1088-9051. doi: 10.1101/gr.1910904. URL http://www.ncbi.nlm.nih.gov/pubmed/15173114.

Zhiqiang Wang, Jing Ma, Chika Miyoshi, Yuxin Li, Makito Sato, Yukino Ogawa, Tingting Lou, Chengyuan Ma, Xue Gao, Chiyu Lee, Tomoyuki Fujiyama, Xiaojie Yang, Shuang Zhou, Noriko Hotta-Hirashima, Daniela Klewe-Nebenius, Aya Ikkyu, Miyo Kakizaki, Satomi Kanno, Liqin Cao, Satoru Takahashi, Junmin Peng, Yonghao Yu, Hiromasa Funato, Masashi Yanagisawa, and Qinghua Liu. Quantitative phosphoproteomic analysis of the molecular substrates of sleep need. *Nature*, 558 (7710):435–439, June 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0218-8. URL http://www.nature.com/articles/s41586-018-0218-8.

Graham H. Diering, Raja S. Nirujogi, Richard H. Roth, Paul F. Worley, Akhilesh Pandey, and Richard L. Huganir. Homer1a drives homeostatic scaling-down of excitatory synapses during sleep. *Science*, 355(6324):511–515, February 2017. ISSN 1095-9203. doi: 10.1126/science.aai8355. URL http://www.ncbi.nlm.nih.gov/pubmed/28154077.

Damian Szklarczyk, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian Von Mering. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, January 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw937.

Sara B. Noya, Franziska Brüning, Tanja Bange, Jan D. Rudolph, Jürgen Cox, Steven A. Brown, Matthias Mann, and Maria S. Robles. Rest-activity cycles drive dynamics of phosphorylation in cortical synapses. *Submitted*, 2018.

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

goes Linux. *Nature methods*, 15:401, June 2018b. ISSN 1548-7105. URL https://www.nature.com/articles/s41592-018-0018-y.

Daniel H. Lackner, Alexia Carré, Paloma M. Guzzardo, Carina Banning, Ramu Mangena, Tom Henley, Sarah Oberndorfer, Bianca V. Gapp, Sebastian M. B. Nijman, Thijn R. Brummelkamp, and Tilmann Bürckstümmer. A generic strategy for CRISPR-Cas9-mediated gene tagging. *Nature Communications*, 6(1):10237, December 2015. ISSN 2041-1723. doi: 10.1038/ncomms10237. URL http://www.nature.com/articles/ncomms10237.

Daichi Kamiyama, Sayaka Sekine, Benjamin Barsi-Rhyne, Jeffrey Hu, Baohui Chen, Luke A. Gilbert, Hiroaki Ishikawa, Manuel D. Leonetti, Wallace F. Marshall, Jonathan S. Weissman, and Bo Huang. Versatile protein tagging in cells with split fluorescent protein. *Nature Communications*, 7:11046, March 2016. ISSN 2041-1723. doi: 10.1038/ncomms11046.

Hyun Woo Rhee, Peng Zou, Namrata D. Udeshi, Jeffrey D. Martell, Vamsi K. Mootha, Steven A. Carr, and Alice Y. Ting. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science*, 339(6125):1328–1331, March 2013. ISSN 1095-9203. doi: 10.1126/science.1230593. URL http://www.ncbi.nlm.nih.gov/pubmed/23371551.

Kyle J. Roux, Dae In Kim, and Brian Burke. BioID: A screen for protein-protein interactions. In *Current Protocols in Protein Science*, volume 74, pages 19231–192314. John Wiley & Sons, Inc., Hoboken, NJ, USA, November 2013. ISBN 0471140864. doi: 10.1002/0471140864.ps1923s74.

Franz Herzog, Abdullah Kahraman, Daniel Boehringer, Raymond Mak, Andreas Bracher, Thomas Walzthoeni, Alexander Leitner, Martin Beck, Franz Ulrich Hartl, Nenad Ban, Lars Malmström, and Ruedi Aebersold. Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science*, 337(6100):1348–1352, September 2012. ISSN 1095-9203. doi: 10.1126/science.1221483.

Fan Liu, Dirk T. S. Rijkers, Harm Post, and Albert J. R. Heck. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nature Methods*, 12(12):1179–1184, April 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3603.

Anders R. Kristensen and Leonard J. Foster. Protein correlation profiling-SILAC to study protein-protein interactions. *Methods in Molecular Biology*, 1188:263–270, 2014. ISSN 1064-3745. doi: 10.1007/978-1-4939-1142-4_18.

Catalina O. Tudor, Karen E. Ross, Gang Li, K. Vijay-Shanker, Cathy H. Wu, and Cecilia N. Arighi. Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. *Database*, 2015, 2015. ISSN 1758-0463. doi: 10.1093/database/bav020. URL `http://www.ncbi.nlm.nih.gov/pubmed/25833953`.

Alice Bossi and Ben Lehner. Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5(1):260, January 2009. ISSN 1744-4292. doi: 10.1038/msb.2009.17. URL `http://www.ncbi.nlm.nih.gov/pubmed/19357639`.

Thorsten Will and Volkhard Helms. PPIXpress: construction of condition-specific protein interaction networks based on transcript expression. *Bioinformatics*, 32(4):571–578, February 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv620.

S. Ohno. So much "junk" DNA in our genome. *Brookhaven Symposia in Biology*, 1972. ISSN 0068-2799. doi: citeulike-article-id:3483106.

Elizabeth Pennisi. ENCODE project writes eulogy for junk DNA. *Science*, 2012. ISSN 1095-9203. doi: 10.1126/science.337.6099.1159.

Acknowledgements

I would like to express my gratitude to the numerous people who made this thesis possible.

Matthias Mann for his supervision and for serving on my thesis advisory committee alongside Professor Caroline Friedel and Jürgen Cox. Your advice was invaluable.

A very special thanks to Jürgen Cox for shaping my work while giving me the freedom to pursue my interests. For, whenever I was stuck, breaking down seemingly impossible problems into a number of simple steps. It was inspirational to learn the ins and outs of computational mass spectrometry from the master himself. Thank you for the opportunities to teach at the various workshops and summer schools, which forced me to learn even more about all the topics surrounding my research.

Charo and Franziska for the fruitful collaboration.

The current and past Cox lab members which made the last years so much more enjoyable.

My gym buddies Shivani, Dan, and Peter for inspiring me to work(out) harder and stick with it. I learned some life lessons with you in the basement.

Susi for all the joy and for believing in me.

My family for having my back. I love you guys!