Complex-centric proteome profiling by SEC-SWATH-MS for the parallel detection of hundreds of protein complexes

Complex-centric proteome profiling by SEC-SWATH-MS for the parallel detection of hundreds of protein complexes

Isabell Bludau[1,]*, Moritz Heusel[1,2,]*, Max Frank[1], George Rosenberger[1,3], Robin Hafen[1], Amir Banaei-Esfahani[1], Audrey van Drogen[1], Ben C. Collins[1,4], Matthias Gstaiger[1,$], Ruedi Aebersold[1,5,$]

[1] Institute of Molecular Systems Biology, Department of Biology, ETH Zurich, Zurich, Switzerland.

[2] Division of Infection Medicine (BMC), Department of Clinical Sciences, Lund University, Lund, Sweden

[3] Columbia University, New York, United States

[4] School of Biological Sciences, Queen's University of Belfast, UK

[5] Faculty of Science, University of Zurich, Zurich, Switzerland.

* These authors contributed equally to this work

[$] Corresponding authors

Abstract

Most catalytic, structural and regulatory functions of the cell are carried out by functional modules, typically complexes containing or consisting of proteins. The composition and abundance of these complexes and the quantitative distribution of specific proteins across different modules is therefore of major significance in basic and translational biology. To date, the systematic detection and quantification of protein complexes has remained technically challenging. The chromatographic separation of native protein complexes followed by the mass spectrometric analysis of the proteins contained in sequential fractions results in potentially thousands of protein elution profiles from which, in principle, the presence of specific complexes can be inferred. However, the de novo inference of protein complexes from such datasets has so far remained limited with regard to selectivity and the retrieval of quantitative information.

We recently developed a variant of this strategy, complex-centric proteome profiling, which extends the concepts of targeted proteomics to the level of native protein complex analysis. The complex-centric workflow consists of size exclusion chromatography (SEC) to fractionate native protein complexes, DIA/SWATH mass spectrometry to precisely quantify the proteins in each SEC fraction based on a consistent set of peptides, and targeted, complex-centric analysis where prior information from generic protein interaction maps is used to detect and quantify protein complexes with high selectivity and statistical error control via the computational framework *CCprofiler*. Complex-centric proteome profiling captures the majority of proteins in complex-assembled state and reveals their organization into hundreds of complexes and complex variants observable in a given cellular state. The protocol is applicable to genetically unaltered tissue cultures and adaptable to primary tissue. At present it requires approximately 8 days of wet-lab work, 15 days of MS measurement time and 7 days of computational analysis.

Introduction

Proteins are major effectors and regulators of biological processes and can elicit or participate in multiple functions depending on their interaction with other proteins and most catalytic, regulatory and structural functions are carried out by protein complexes. Therefore, it is of central and general interest in basic and translational biology to identify protein complexes in biological samples and to detect changes in their composition and/or abundance across different samples and states. Over the last decades, different strategies to systematically study protein-protein interactions (PPIs) and protein complexes have been developed. With increasing technical abilities in the area of mass spectrometry based proteomics, high-throughput methodologies for detecting PPIs by affinity-purification coupled to mass spectrometry (AP-MS) have emerged as gold-standard technique for mapping large PPI interaction networks [1–3]. However, the need to perform multiple reciprocal pull-downs, requiring either genetic engineering or availability of multiple specific antibodies, makes the AP-MS approach limited in its capability to comprehensively study the protein interactome across multiple conditions and to reliably detect in parallel concurrent changes in multiple complexes in a set of samples. Other techniques, e.g. BioID [4], are based on proximity labeling and obtain information about proteins that are in close spatial proximity inside the cell but that do not necessarily interact directly. Another recently developed technique couples native protein co-fractionation of complexes with mass spectrometry followed by correlative analysis of chromatographic profiles to infer protein-protein interactions (CoFrac-MS or PCP-MS) [5–12]. Here, a single fractionation experiment in principle captures information on thousands of PPIs and hundreds of protein complexes in parallel and substantially increases the throughput for screening protein complexes.

Previous applications of CoFrac-MS strategies already provided interesting novel insights into the protein complex landscape across different organisms [10] and molecular perturbations [13]. They also showed that the CoFrac-MS approach faces the general limitation that the cumulative number of proteins detected across the fractions exceeds the number of fractions and by almost two orders of magnitude. Further, the methods remained limited with regard to the resolution of chromatographic separations, the accuracy and consistency of mass spectrometric protein measurements as well as the specificity in the inference of protein-protein interactions and the composition of protein complexes from the highly convoluted data [7,12,14].

With the goal to address these prevailing limitations and to assess proteome organization more quantitatively and precisely, we developed the complex-centric proteome profiling strategy that builds on SEC-based protein complex fractionation with increased chromatographic resolution, optimized peptide and protein quantification by data-independent acquisition mass spectrometry (DIA/SWATH-MS) and the targeted detection of protein complexes at controlled error-rates based on queries of putative protein complexes in complex-centric analysis [15]. The core difference to other approaches lies in the concept of complex-centric analysis which essentially changes the data analysis strategy of CoFrac-MS data from a discovery-based strategy to a multiple hypothesis testing based strategy that uses prior information of protein complexes to suggest stable modules. In essence, the complex-centric strategy is an extension of the targeted proteomics rationale (Box 1) from the level of protein analysis to the level of protein complex analysis. The targeted query of protein complexes in the high quality and consistent co-fractionation data generated by DIA/SWATH-MS improves selectivity compared to other, discovery based data analysis approaches that generally focus on the detection of pairwise interactions from which complex compositions are then estimated (interaction-centric, [14,16]). In addition to accurate detection, complex-centric analysis also extracts quantitative information on the composition and abundance of the cellular module(s) in which each detected protein participates. Complex-centric proteome profiling thus supports the parallel detection and quantification of hundreds of protein complexes at unprecedented resolution and bears significant potential to discover novel aspects of modular proteome function in diverse biological processes.

Overview of the protocol

This protocol article provides step-by-step instructions to profile the higher order complex assembly state of a given proteome via the complex-centric SEC-SWATH-MS workflow. The workflow consists of three main modules: (1) Sample preparation via extraction of protein complexes under native conditions from a biological sample and their fractionation by size exclusion chromatography (SEC); (2) Data acquisition by bottom-up proteomics analysis of all collected fractions by data-independent acquisition mass spectrometry and targeted, peptide-centric analysis (SWATH-MS); and (3) Computational inference of proteome assembly state and detection of specific protein complexes by targeted, complex-centric analysis within the R software package *CCprofiler*.

The protocol takes as input a biological sample, exemplified by 7e7 HEK293 cells, as well as prior information on the tested protein interactions and/or complexes formed within the proteome of interest. The protocol produces as output (1) a quantitative assessment of the global proteome assembly state of the proteome analyzed, (2) a quantitative assessment indicating how each protein partitions into a certain number of SEC-resolvable, distinct protein assemblies, and (3) a quantitative assessment of the interactions and protein complexes in the given proteome in the biological state tested. Figure 1 schematically summarizes the overall complex-centric proteome profiling workflow by SEC-SWATH-MS.

The key requirements to successfully perform a SEC-SWATH-MS experiment and to analyze the resulting data are (1) the availability of biological specimens from which intact native complexes can be isolated, (2) HPLC and DIA-enabled mass spectrometric equipment and (3) computing infrastructure for data storage, peptide-centric SWATH-MS data analysis as well as complex-centric SEC-MS data analysis in the R environment. The built-to-task algorithms for complex-centric analysis are implemented in the R package *CCprofiler*. Familiarity with the R programming language is not required for the use of the canonical complex-centric workflow but highly recommended if deviations from the canonical workflow as presented herein are required.

Applications of the method

The key feature of complex-centric proteome profiling by SEC-SWATH-MS is its ability to assign separated proteins to hundreds of protein complexes in parallel from the same sample and to determine the relative distribution of protein mass across different assembly states. These analyses provide results at controlled error-rate and at sub-complex resolution. Preliminary results show that the analysis strategy can be extended to quantitatively compare protein complex abundance and composition and the distribution of proteins into alternative assembly states across different biological conditions, which is the focus of ongoing work.

Compared to other co-fractionation methods, the complex-centric workflow supports the rapid and error-controlled detection of hundreds of protein complexes from a single dimension of chromatographic fractionation, supported by key improvements in quantitative profiling and data analysis. The workflow extends the targeted proteomics rationale as initially implemented by selected reaction monitoring (SRM) and more recently by DIA/SWATH-MS towards the detection of protein complexes from CoFrac-MS data in targeted queries of database-curated complexes or other sources providing prior information about the protein complexes to be tested. We could demonstrate that the targeted analysis rationale improves the overall selectivity of detecting reference complexes to a level comparative to that of a state-of-the-art CoFrac-MS workflow based on the analysis of a 12-fold higher number of fractions generated by multidimensional biochemical separations and binary interaction-centric data processing [9,10,15]. In addition to indicating complex composition, the complex-centric workflow also provides insights into the abundance of the subunits (complex stoichiometry) as well as the relative distribution of specific proteins across the detectable complexes in which they participate. These types of information are expected to closely correlate with the functional state the respective protein modules, based on the principle that biochemical activities of a protein complex substantially depend on its subunit composition, topology and overall structure and that alterations of these parameters also alter activity and function [17].

While the first exemplary application of the protocol was shown for the HEK293 cell line [15], the complex-centric methodology can be readily adapted to different sample types that are compatible with mild lysis and the extraction of sufficient amounts of native protein complexes for SEC-SWATH-MS analysis (at least 1 mg of total protein). The complex-centric proteome profiling strategy can thus serve to assess proteome organization and assembly states of specific protein complexes across a wide range of experimental model systems and perturbations, including clinical specimen. Besides studying essential biological processes such as e.g. the impact of growth factor treatment or cell differentiation, it can be envisaged that SEC-SWATH-MS can also be employed towards the characterization of drug mode of action with respect to protein complex assembly states or in fact for biomarker discovery. For example, complex-centric profiling might capture off-target effects that do not affect the level of transcripts or proteins but that alter cellular functions by altering the assembly state of specific protein complexes. Application of the complex-centric proteome profiling workflow to different experimental systems or across perturbations bears profound potential to provide novel insights into the interplay between proteome organization and function. A core focus of future work will therefore be to further improve method throughput to enable larger comparative or longitudinal studies, for example by high throughput on-line C-18 liquid chromatography [18]. Further, we envisage the extension to additional fractionation techniques and the extension of the computational framework CCprofiler to quantitatively compare protein complex assemblies across conditions both on the level of protein complex assembly states and protein complex composition.

The overall performance of the presented workflow is tightly linked to chromatographic resolution, the accuracy of mass spectrometric readout and the selectivity of complex-centric analysis. We thus recommend applying all three modules of the workflow as presented herein. However, depending on the specific biological question at hand and the available technology and expertise in a research laboratory, we anticipate that individual modules of the workflow can be replaced by alternative approaches or adapted to specific conditions. For example, protein complexes could be fractionated by other techniques than SEC, such as density gradient centrifugation, blue native poly-acrylamide gel electrophoresis (BN-PAGE), or ion-exchange chromatography (IEX), as has been employed in other studies [9,10,19]. It should be noted, however, that the fractionation technique employed will determine key attributes such as resolution and observability of certain protein complexes that may impair workflow performance, especially in cases where no specific size information is obtained as is the case in IEX. Available size estimates of the detected protein complexes improve the selectivity of the canonical SEC-SWATH-MS workflow. Furthermore, bottom-up proteomic quantification can alternatively be achieved by conventional data-dependent acquisition (DDA) in combination with MS1-based or spectral counting-based quantification, in contrast to DIA/SWATH-MS and MS2-based quantification as is suggested in this protocol. However, depending on the MS instrument and acquisition setup, this can result in a reduced quantification precision and consistency, thereby leading to less sensitive protein complex detection (Figure 2C in [15]). Finally, if the discovery of novel protein complexes is of interest, the data analyzed with regard to quantitative assembly states of previously known protein complexes via *CCprofiler* can alternatively or in addition be employed to predict novel complexes via analysis tools such as PrInCe [14] or EPIC [20]. This might be especially useful in cases where no or only limited prior knowledge on protein complexes is available for the organism analyzed.

Experimental design

Complex-centric proteome profiling by SEC-SWATH-MS was designed to maximize the resolution and depth of biological insights obtainable from a single protein co-fractionation experiment. This is achieved by extending concepts from the analysis of peptides via targeted proteomics to the level of protein complexes, thus implementing a targeted data analysis strategy. The workflow extracts quantitative information on the observed assembly states of predefined protein complexes with high selectivity. The workflow consists of three main modules, each of which was optimized to address the prevailing limitations of previous co-fractionation workflows.

(1) Extraction of protein complexes from a biological sample and fractionation by size exclusion chromatography

To minimize complex disassembly under the diluted conditions present after cell or tissue lysis, all processing steps up to the collection of fractions are carried out rapidly, at minimal dilution and below 4°C. To improve on protein complex resolution and to produce strongly correlated, sharp peak signals for the detected proteins, SEC is performed with high-resolution stationary phase material that offers an optimal tradeoff between resolution and fractionation range. A preferred column material combines 3 µm particle size and 500 Å pore diameters.

Protein complexes are isolated using a mild detergent, nonidet-P40, that has been shown to support the extraction and maintenance of the integrity even of relatively labile protein complexes [21–23]. To minimize the artifactual disruption of native protein complexes cells are snap-frozen in liquid nitrogen prior to immediate processing. The cells are lysed by freeze-thawing into a low volume of 0.5% detergent-containing lysis buffer supplemented with protease- and phosphatase inhibitors. These precautions are taken to minimize artifactual proteome rearrangements that do not reflect the true biological state of the system. After proteome extraction, lysates are cleared by ultracentrifugation. This step is pivotal to remove cellular components interfering with successful SWATH-MS analysis. To reduce detergent levels and to concentrate the sample for SEC fractionation, the buffer is exchanged over a 30 kDa molecular weight cut-off membrane in multiple dilution steps to minimize dilution-induced complex dissociation. To monitor column performance and calibrate the apparent molecular weight scale per each fraction, a 5-protein standard sample is analyzed prior to and after the preparative protein complex fractionations required for an experiment or study. The elution profiles of standard proteins are further used to determine the fraction collection scheme. We generally aim at collecting 7-8 independent MS measurements (fractions) across each eluting peak, a measure that has proven optimal in targeted proteomics applications aimed at the accurate quantification of peptides based on fragment ion chromatogram peak groups along peptide elution from on-line C-18 chromatography [24,25].

An aliquot of the unfractionated sample should be included for SWATH/DIA LC-MS/MS analysis to obtain a 'master' sample and SWATH-MS dataset in which most if not all signals observable across the individual fractions are represented. This dataset is used to optimize peptide-centric analysis, specifically to align parameters for statistical scoring of peptide fragment ion peak group signals across the full SEC-SWATH-MS study.

(2) Bottom-up proteomics analysis and peptide-centric analysis (SWATH-MS)

To improve the accuracy and consistency of mass spectrometric peptide signals along the SE chromatographic axis, data-independent acquisition-based SWATH mass spectrometry is employed. Prior to sample acquisition proteins contained in each collected SEC fraction are denatured by heating in 1 % sodium de-oxycholate, reduced, alkylated and trypsinized overnight. Compared to a range of alternative sample workup regimens tested, the protocol described below provided most robust results. The tryptic peptides are cleaned-up by binding to a C18-resin and subsequent one step elution. They are then dried and resuspended in nanoLC buffer A, supplemented with internal standard peptides to align peptide retention times in C18 chromatography, essentially as described [26]. The inclusion of the reference peptides is especially important for peptide-centric, spectral library-based data analysis (see below).

To maintain the ability to quantitatively compare peptide/protein abundances across individual fractions of the recorded SEC-SWATH-MS maps, we recommend the injection of equal volumes per each SEC fraction for analysis by SWATH-MS. We recommend to determine the injection volume for a specific SEC separation based on test injections of the first two SEC fractions that produce highest absorbance at 280nm in UV/Vis profiling during SEC fractionation. The injection volume is then selected to maximize the amount of peptides injected per MS run without exceeding the upper sample limit tolerated by the mass spectrometer (Figure 2A). This upper limit is determined by detector saturation. For example, for a 5600+ instrument operated in SWATH 64vw mode the maximal total ion current (TIC) is 1e8 (here fraction 55, also see Figure 2B). As a consequence of keeping the volume injected from each SEC fraction constant, fractions containing smaller amounts of total protein will generate SWATH maps that top out significantly below the maximally tolerated ion current. The identity of the most highly concentrated fractions and optimal injection volume may depend on sample type, SEC setup and MS platform used. We recommend including an aliquot of the unfractionated sample in the final acquisition queue.

While other MS platforms and MS acquisition schemes may in principle be also compatible with complex-centric profiling, we advise against employing classical DDA-MS acquisition and quantification because the performance of complex-centric profiling and specifically the selectivity of data analysis in *CCprofiler* strongly depend on the improved quantitative accuracy achievable by DIA/SWATH-MS [15,27]. The workflow described was successfully implemented using the ABSciex 5600+ or 6600+ MS platforms in combination with the 64 variable window acquisition scheme (also see Tables 1, 2 and 3) [15,27]. Other quantitative DIA workflows as implemented on alternative mass spectrometric platforms may also produce data of sufficient quality for effective complex-centric proteome profiling [28–30]. We recommend acquisition of SWATH-MS maps along the SEC fractions in 120 min gradients, yet recent studies have shown promising results when employing shorter (e.g. 60 min) nanoLC or super-short fastLC gradients (20min, [18]).

The interpretation of DIA/SWATH-MS data is most effective in conjunction with a reference spectral library that contains information about peptide elution, MS fragmentation and other pertinent properties. These parameters then support the detection and quantification of the respective peptides by targeted data analysis. Whereas samples from less intensely studied species may require the generation of a customized spectral library from DDA data acquired on the same sample [26], for more extensively studies species, including human, mouse, yeast and zebrafish, large scale libraries are available via http://www.swathatlas.org/ [31–33]. For applications involving human samples, we recommend using the combined human assay library (CAL) that combines peptide query parameters from over 300 DDA-MS injections across diverse human tissue types [31]. The CAL also contains spectral information from our complex-centric analysis of the HEK293 proteome [15] and is thus expected to be sufficiently representative also for future studies of human (cell lines) via SEC-SWATH-MS. In the course of developing the SEC-SWATH-MS methodology, the use of a sample-specific vs. the combined human assay library was evaluated. The results showed benefits of the CAL with respect to increased sequence coverage and increased number of peptides per protein [34]. If a given research question depends on a sample-specific spectral library, we recommend DDA acquisition side-by-side with the DIA/SWATH-MS measurements and library construction as described previously [26].

Once all DIA/SWATH-MS data is acquired, data are converted to mzXML format for open-source peptide-centric analysis via the OpenSWATH workflow which consists of OpenSWATH, PyProphet and TRIC analysis [35–40]

(3) Analysis of the global proteome assembly state and specific protein complexes by targeted, complex-centric analysis via *CCprofiler*

The overall goal of the complex-centric workflow is the inference of the global proteome assembly state of the sample tested and the detection of specific protein complexes from the measured SEC protein profiles. We generated *CCprofiler*, an algorithm and computational framework to support these analyses. After initial data preparation and import, *CCprofiler* performs three main steps: (i) data pre-processing and quality control, (ii) protein-centric analysis and (iii) complex-centric analysis. In the following, the rationale behind each of these three steps is explained.

Data preparation and import

Prior to analysis in *CCprofiler*, all necessary input data first needs to be formatted and imported into R according to the *CCprofiler* guidelines (also see Table 1). The main input for *CCprofiler* is the quantitative peptide-level data, which needs to be imported into a traces object. This is a customized data format of the *CCprofiler* package. Traces objects can store information about quantitative peptide or protein profiles and can further include additional annotations of the measured peptides, proteins and fractions. For direct data import of the TRIC output, the importFromOpenSWATH function can be used. It removes non-proteotypic evidence and sums precursor signals per peptide to generate peptide-level quantitative information. If an alternative DIA analysis platform is selected, non-proteotypic peptides should be removed manually if necessary and precursor signals should be summarized to unique peptide quantitative information. The thereby generated quantitative peptide matrices can be also imported into traces objects via the importPCPdata function. In addition to (1) the quantitative peptide-level information, *CCprofiler* requires (2) a fraction annotation table that maps each MS run to a given fraction, (3) a molecular weight (MW) calibration table generated by measuring the apex fractions of an external standard set of reference proteins fractionated on the same SEC setup that is used to establish a log-liner-relationship between the SEC fractions and their the apparent molecular weight (Figure 3A), (4) a trace annotation table containing information from UniProt (https://www.uniprot.org/) that is used to annotate proteins with according gene names and monomeric MWs, and finally (5) prior protein connectivity information in form of defined protein complexes, e.g. annotated in CORUM [41], or binary interaction networks, such as StringDB [42,43] or BioPlex [1,2]. A summary of the required input data and format for a successful *CCprofiler* analysis are summarized in Table 4.

Data pre-processing and quality control

To achieve optimal data quality and sensitivity in complex-centric profiling, *CCprofiler* includes several functions to increase data completeness and to filter peptides for detection and quantification consistency.

Once a traces object is imported, missing values can be detected based on a user defined criterion. In most proteomics pipelines, zero intensity values indicate either that the signal is "missing at random", i.e. no detection due to technical reasons such as interferences from other peptides, or "missing not at random" i.e. no detection due to cellular concentrations below the detection limit. We suggest that a zero value is considered as "missing at random" in case a quantitative (non-zero) signal has been detected in both, the two fractions preceding or following the fraction in question. The detected "missing at random" values are subsequently imputed by a spline fit across the fractionation dimension. In our originally published analysis workflow [15] we did not perform a missing value imputation. Generally, when applying a SWATH-MS based workflow, only few missing values are expected. However, sample loss, e.g. of an entire SEC fraction, sometimes cannot be avoided, thus generating missing values for transparent technical reasons. To still enable a robust and sensitive analysis with *CCprofiler* in such cases, we implemented the missing value detection and imputation approach. While this does not make a large difference in the HEK293 SEC-SWATH-MS dataset presented in this paper, we specifically recommend this pre-processing step for datasets generated by DDA or if single fractions were lost during sample processing.

In a next step, peptides that have never been detected in more than N consecutive fractions, here N=2, are removed from the traces object. This effectively removes false positive peptide detections from the dataset. Finally, we leverage the idea that multiple peptides originating from the same protein should display highly similar quantitative elution profiles along the chromatographic dimension, given that the proteins are presumably intact during separation and are only cleaved into peptides for LC-MS/MS analysis thereafter. For each peptide, the average pairwise correlation with the quantitative traces of its sibling peptides, i.e. peptides derived from the same protein, is calculated (Figure 3B). Peptides below a minimum average sibling peptide correlation (SPC) cutoff can subsequently be removed. The rational for this step is that outlier peptides as well as proteins with very heterogeneous quantitative peptide traces are excluded from further analysis. Given that decoy proteins are maintained in the upstream analysis, *CCprofiler* can automatically determine an SPC cutoff at which a user-defined criterion on the maximally acceptable global protein-level FDR is satisfied (Figure 3C). FDR estimation can be fine-tuned by providing a prior on protein detectability in form of a fraction of false targets (FFT). A conservative FFT can be estimated from the protein-level PyProphet scoring of the unfractionad SEC input sample. This is conservative, because we expect to see cumulatively more proteins in the SEC fractions than in the single unfractionated input sample. Alternatively to the automated SPC cutoff estimation, a user-provided SPC cutoff can be used for peptide filtering. Figure 3E shows the peptide-level quantitative profiles of the Proteasome subunit alpha type-1 after pre-processing and filtering.

The peptide-level traces can subsequently be used for protein quantification. The protein level profiles are inferred from summing the top N (we recommend N = 2) most-abundant and well-correlated peptides of a given protein. Only peptides that map uniquely to the protein of interest and in the context of the full sequence database (proteotypic peptides) are used in the analysis.

The protein-level profiles can then be used to estimate the overall complex assembly state observed in the sample as a quality control to ensure the successful extraction and profiling of largely intact complexes (Figure 3D). The protein-level profiles are further the input for complex-centric signal detection.

Protein-centric signal detection of protein assembly states

Protein-centric analysis aims to evaluate the number of distinct assembly states each protein is observed in. We define as an assembly state a distinct SEC peak in which the protein in question is confidently detected. The analysis thus detects peptide co-elution peak groups along the chromatographic dimension. Each detected peak (‘protein feature’) represents the protein in a specific assembly state, i.e. monomeric or bound to different protein complexes, as inferred from the proteins' monomeric MW and external size calibration of the SEC fractionation. This analysis yields a fine-grained view of individual assembly states of each protein but also enables more global assessments of the overall degree of higher order assembly observed from the biological sample.

The optimal parameters for the peak detection algorithm in *CCprofiler* depend on the chromatographic resolution and the quantitative accuracy of a given dataset. Therefore, *CCprofiler* features functions to automatically screen and identify optimal parameters based on a parameter grid search (Box 2 and Figure 4A). The grid search screens different parameter combinations for their sensitivity to detect protein elution signals from sibling peptides (derived from the same protein) as opposed to nonsense signals among peptides sampled randomly from different proteins across the dataset. In our previously published analysis workflow [15], parameter optimization was performed based on the detection of protein complex signals from protein traces. In contrast, we here present and suggest parameter optimization based on protein-level peak detection from peptide traces, because it is independent of the detectability of a certain set of protein complexes in the sample and because the chromatographic and quantitative properties are specific to the analyzed dataset and not the level of analysis (complex or protein elution).

Scoring and FDR estimation of the detected protein features is performed based on a target decoy strategy. Decoys are generated automatically, by randomly shuffling peptide-to-protein assignments. The co-elution scores calculated by *CCprofiler* (for more details also see information available in the appendix of the original publication [15]) are converted into empirical p-values and used for q-value estimation [44]. To ensure that the decoy based FDR estimation approach works correctly, the quantitative data needs to be of sufficient quality and a sufficient number of proteins need to be evaluated in parallel. We suggest that at least 500 proteins with high-quality SEC patterns that pass upstream consecutive identification and sibling peptide correlation-based filters (see Step 37iii) should be used. To assess whether the FDR estimation worked correctly, it is crucial to manually inspect the p-value density histogram generated by *CCprofiler* (Figure 4B/C). There should be a high peak close to zero and a uniform distribution across all other p-values (also see troubleshooting). Figure 4E shows the three unique elution signals of the Proteasome subunit alpha type-1detected by protein-centric analysis.

The confidently detected protein features each represent a unique protein assembly state, which allows the inference of several interesting biological conclusions. First, the distribution of protein features into signals likely representing monomeric or assembled forms of the protein can be used to draw conclusions about the overall assembly state of the cellular system tested (pie chart in Figure 4C). Second, the distribution of proteins into a single or multiple distinct assembly states can reveal potential moonlighting of a protein across distinct protein assemblies in which it may assume alternative functional roles (bar chart in Figure 4C). Protein feature signals do, however, not yet assess the exact interaction partners involved and their quantitative distribution across specific protein complexes, which is a question that is assessed in the subsequent complex-centric signal detection module.

Complex-centric detection of protein complex signals

The heart of the *CCprofiler* software and the presented protocol is the complex-centric detection of protein complex signals. Here, prior protein connectivity information is used to query the SEC-SWATH-MS data for evidence of specific protein complexes or sub-complexes thereof in the biological sample. In analogy to peptide- and protein-centric analysis, complex-centric analysis detects protein co-elution peak groups along the chromatographic dimension based on *a priori* defined protein complex queries and test the hypothesis that the query complex be present in the sample. A single protein complex query frequently results in the detection of multiple distinct subunit co-elution signals (‘protein complex features’). Each of these signals corresponds to a different (sub-)version of the queried protein complex with distinct composition and/or stoichiometry, eluting at a distinct elution time. Depending on the completeness of prior information and MS observability (Figure 5A), targeted, complex-centric analysis may or may not capture all involved protein subunits. In contrast to other analysis strategies of co-fractionation MS datasets that aim to predict new complexes [16,20], the complex-centric workflow aims to confidently detect *a priori* defined protein complexes and variants thereof and to quantify protein subunit distribution across these in the given biological state.

Prior protein connectivity information can be provided either directly in the form of the concrete composition of query complexes ('complex queries') or in form of pair-wise protein-protein interaction (PPI) networks. An exemplary set of target complex queries are the complexes represented in CORUM [41]. If a PPI network is provided, it first needs to be partitioned into smaller sets of interacting proteins, thereby generating concrete protein complex queries. This is achieved by simplistic network partitioning, selecting for each protein its immediate neighbors (degree = 1). Any redundant, fully overlapping complex queries are removed such that per unique query complex one decoy query complex is generated, and the assumptions for the error control strategy are met. However, we suggest to keep protein complex subset queries, such as e.g. a first complex query including subunits A, B and C and a second complex query including subunits A, B, C, D and E. The reason for keeping also the sub-complex query ABC is that the co-elution signal A-B-C will be scored preferably in the *CCprofiler* co-elution score when compared to the same signal scored as part of a larger complex query including subunits D and E. This is because scores are calculated in dependence of the number of proteins per query and the number of proteins observed co-eluting. Maintaining the subset hypothesis A-B-C will therefore increase the chance to successfully recover and quantify the A-B-C complex signal, particularly in the situation where subunits D and/or E do not partake in the observed protein complex and co-elution signal (for more details also see the appendix on 'Coelution score calculation and statistical FDR control' in [15]). For cases in which both, the A-B-C and A-B-C-D-E complex query point to the same protein complex signal in the SEC dimension, these detected features will finally be collapsed into one unique signal after the initial peak detection step (see below).

To enable an automated error-estimation of the complex-centric feature finding, *CCprofiler* employs a decoy based FDR estimation strategy. A decoy complex query is generated for each target, thus resulting in the same size distribution of protein complex subunits for both targets and decoys. Decoys are generated by randomly assigning proteins to a specific decoy complex query. To ensure that the randomized protein-to-decoy-complex associations do not contain true interactions, we exclude direct interaction partners present in the target protein complex queries. This is achieved by selecting a minimum pairwise network distance N, here N=2. The target-decoy FDR control strategy of *CCprofiler* depends on a minimal number of target (and decoy) queries. Reasons are that (a) a sufficiently large binary interaction network needs to be available to randomly generate decoy complex queries and (b) decoy-based FDR estimation is only appropriate for a representative number of detected target and decoy queries in order to ensure a stable p- and q-value estimation. We suggest a minimum number of 1000 protein complex targets and decoys.

After the target and decoy protein complex queries are defined, complex-centric feature finding is performed. Here, protein traces are queried for the presence of local co-elution signals of the subunits specified in each of the target and decoy queries. Co-elution scores are calculated based on both the local correlation of the detected subunits across the SEC fractions as well as the fraction of correlating subunits relative to the total number of queried subunits. For the estimation and control of FDR, the most-complete co-elution signal for each protein complex query, i.e. the peak with the highest number of co-eluting protein subunits, is selected and used to convert co-elution scores into empirical p- and q-values [44]. To confirm the correct operation of the error estimation and the fidelity of the results, it is important to manually inspect the p-value density histogram generated by *CCprofiler* (Figure 5B/C). There should be a high peak close to zero and a uniform distribution across all other p-values (also see troubleshooting). To interrogate potential protein complex assembly intermediates or other protein complex variants that are observable from a set of query subunit profiles additional to the best scoring peak group, secondary peak groups of each complex query are subsequently appended to the FDR filtered list of detected protein complex signals. For these assignments a less strict minimum local peak-correlation cutoff is selected manually specifically for the respective dataset.

The list of confidently detected protein complex features contains information not only about the presence and abundance of individual protein complexes, but further entails information about proteome modularity such as protein complex assembly intermediates and subunit stoichiometry (also see Figure 5D/E and anticipated results section).

Up to this point in the complex-centric analysis workflow, each protein complex signal detected by *CCprofiler* is directly linked to one specific protein complex query provided by the prior protein connectivity information. However, some of the subunits in each complex query might overlap with other complex queries. This redundancy in complex queries, in combination with the possibility to observe subsets/sub-complexes of these queries can result in the reporting of multiple protein complex signals based on only one piece of experimental evidence (co-elution signal). For example, complex query A consists of subunits WXYZ and complex query B consists of subunits VXYZ. If a co-elution signal among XYZ is detected in the data, it will, until this point, be reported for both complex query A and B. Therefore, in order to retrieve truly unique signals, the co-elution signals need to be collapsed based on a strategy that considers (i) subunit composition and (ii) position in the chromatographic dimension, i.e. apparent molecular weight of the protein complex. In the example above, signal collapsing will merge the redundant report of the XYZ signal derived from the two partially overlapping complex queries A and B into one unique reported protein complex signal (for more details also see [15]). In case that multiple signals with overlapping components are detected at the same elution fraction, unique composite signals will be generated, based on user-defined parameters on subunit overlap and proximity in elution fraction (e.g. signals containing the subunits XYZ, YZK and YZF all with apex fraction 32 will result in one unique reported signal XYZKF). It is important to note that the FDR estimation strategy in *CCprofiler* operates at the level of protein complex queries and does not propagate to the level of collapsed protein complex signals and the precise error estimations among those collapsed results remain de facto unknown and the subject of future work.

Limitations

There are several limitations associated with the presented protocol. First, the workflow is optimized towards the analysis of soluble, cytosolic protein complexes that are extractable under native conditions and that remain stable during the multiple steps of the protocol through SEC separation.

Conceptually, the targeted, complex-centric analysis strategy is not designed to identify any novel protein complexes. Rather, it focuses on the detection and quantification of protein complexes annotated in public protein interaction maps [1,2,41–43]. Therefore, the workflow is limited by the availability and coverage of such prior protein connectivity information. While the quality of the chosen prior interaction network is naturally influencing the results obtainable by complex-centric analysis, the co-elution signal detection step and the FDR model in *CCprofiler* provide a good strategy to reduce the negative impact of false or inaccurate protein complex assignments represented in the prior network. Overall, the targeted approach is more sensitive and selective for protein complex queries with higher numbers of (>3) protein subunits, because random co-elution of these subunits becomes less likely [15] compared to protein complexes consisting of a lower number of subunits.

One critical consideration is the level of redundancy in the prior protein connectivity information used as input for complex-centric analysis. While redundancy, e.g. in the form of larger and smaller protein sub-complexes, can significantly boost sensitivity and protein complex recovery, it also bears the potential to recover the same protein complex signal multiple times from the perspective of different, partially redundant queries. While the signal collapsing strategy in *CCprofiler* in principle removes such redundant complex assignments, FDR control does not propagate throughout this step of the workflow and the results should be treated with more caution.

Finally, the protocol has the same caveats as most large-scale data analysis approaches. Since strict FDR control is necessary to warrant overall high quality results, weaker, yet interesting signals might be missed. Therefore, it might still be advisable to manually inspect chromatograms of specific candidate proteins and protein complexes of interest. In case you can clearly determine the signals manually, an adjustment of the selected parameters for the *CCprofiler* analysis might be necessary in order to obtain optimal results.

Materials

Reagents and consumables

- HEK293 cell line (American Type Culture Collection Cat. No. CRL-1573). Unpublished work has shown that the protocol is applicable also to other cell and tissue types.
- Cell culture dishes 15 cm (Corning, Sigma-Aldrich Cat. No. CLS430599-60EA).
- BCA protein assay kit (Pierce, Fisher Scientific Cat. No. 23225)
- Injection vials 32×11mm (BGB Analytik Cat. No. PPSV0903K&090304)
- 96-DeepWell Plates (Nunc, Fisher Scientific Cat. No. 260251)
- 96-well MacroSpin Plates C-18 (Harvard Apparatus Cat. No. 74-5617)
- 96-well plate adhesive aluminum seals (VWR Cat. no. 60941-112).
- DMEM Thermo Fisher (Gibco, Fisher Scientific Cat. No. 670116)
- Penicillin-Streptomycin-Glutamine 100× (Gibco, Fisher Scientific Cat. No. 10378016)
- Fetal bovine serum (Gibco, Fisher Scientific Cat. No. 26140079)
- Trypsin-EDTA 1×
- Base and SEC buffer:
- N-2-Hydroxyethylpiperazine-N-2-Ethane Sulfonic Acid (HEPES) (Gibco, Sigma-Aldrich Cat. No. 11344041)
- Sodium Chloride (Sigma-Aldrich Cat. No. S7653)
- Lysis buffer (Lysis):
- Sodium Fluoride (Sigma-Aldrich Cat. No. S6776)
- NP-40 detergent (Nonidet P-40, Sigma-Aldrich IGEPAL-630, Cat. No. I8896).
- Sodium Pervanadate Na_3VO_4 (Sigma-Aldrich Cat. No. S6508).
- PMSF (Sigma-Aldrich Cat. No. 78830)
- Protease inhibitor cocktail (Sigma-Aldrich Cat. No. P8340)
- SEC standard proteins (Phenomenex Cat. No. AL0-3042)
- Tris(2-carboxyethyl)phosphine hydrochloride (TCEP, Sigma-Aldrich Cat. No. 75259)
- Iodoacetamide (Sigma-Aldrich Cat. No. I6125
- Trypsin (Promega Sequencing grade, Cat. No. V5111)
- Trifluoroacetic acid (Sigma-Aldrich Ca. No. T6508).
- Retention time normalization kit (iRT kit, Biognosys AG Cat. No. Ki-3002-1).
- Acetonitrile, gradient grade (Sigma-Aldrich Cat. No. 34851)

Equipment

- Sterile cell culture work bench.
- Humidified incubator, 37˚C, 5% CO_2.
- Light microscope for inspection of cell lines.
- 96-well plate incubator.
- Tabletop Microcentrifuge (Eppendorf Centrifuge 5418 or similar, capable to spin at 16,900×g)
- Tabletop centrifuge (Eppendorf Centrifuge 5810R or similar capable to spin at 3,220×g)
- Ultracentrifuge (Beckman Coulter Optima TLX or similar, target: 100,000×g)
- Ultracentrifuge Rotor (Beckman Coulter TLA-120.2 10×2ml, or equivalent)
- Thick-wall Polycarbonate tubes for TLA-120.2 (1ml, Beckman Coulter Cat. No. 343778).
- (Alternative: Rotor TLA-120.1 14×0.5ml with Tubes 0.5 ml 343776).
- Amicon Ultra-4 Centrifugal Filter Units (Sigma-Aldrich UFC803008).
- HPLC system with UV detector, e.g. Agilent 1100 series. Backpressures of ca. 100 bar are typical (500 ul/min).

- Yarra 3um SEC-4000 column (300×7.8 mm, pore size 500 Å, particle size 3 µm, Phenomenex, Cat. No. 00H-4514-K0)
- SecurityGuard column guard cartridge holder (Phenomenex, Cat. No. KJ0-4282)
- Guard column cartridges (Phenomenex Cat. No. AJ0-4489)
- Water bath (VWR Cat. No. 97055-806).
- µl-volume UV-Vis Spectrophotometer (e.g. NanoDrop ND-1000, Thermo Fisher Scientific)
- LC-MS/MS System of nano-LC and DIA/SWATH-MS-enabled mass spectrometer (Eksigent AS-2/1Dplus and AB SCIEX TripleTOF 5600+)
- PicoFrit self-pack columns and emitters (New Objective, Cat. No. PF360-75-10-CE-5)
- Magic C18 Aq resin (3 µm, 200-Å, Michrom H254)
- Windows computer for file conversion with following software
  - ProteoWizard (http://proteowizard.sourceforge.net/)
- Workstation or server computer (any operating system) with ca. 300 gb disk space, 16 gb RAM and ≥ 8 threads/CPU cores with the following software installed.
  - Docker (e.g. https://docs.docker.com/docker-for-windows/)
  - OpenSwath docker pipeline (See http://openswath.org/en/latest/docs/docker.html)
  - R (≥v3.60, https://cran.r-project.org/bin/windows/base/) with packages devtools, data.table, ggplot2 and *CCprofiler* (See setup))

Reagents setup

SEC buffer. The SEC mobile phase is 50 mM HEPES pH 7.5, 150 mM NaCl. 10× stock solutions can be stored at 4°C for up to 8 weeks. Per experiment, prepare 1000ml SEC buffer per experiment to accommodate for system and column equilibration. Use milli-Q water. Before use remove particles by 0.22µm-filtration and store at 4°C.

HNN buffer. HNN Buffer is equivalent to the SEC mobile phase, supplemented with 50 mM NaF for phosphatase inhibition. After 0.22µm-filtration HNN buffer is stable at 4˚C for 8 weeks.

HNN Lysis buffer. HNN Lysis buffer is 0.5% (vol/vol) NP40, 50 mM HEPES, pH 7.5, 150 mM NaCl, 50 mM NaF, 200 µM Na_3VO_4, 1 mM PMSF, and 1× protease inhibitor cocktail. The buffer is to be prepared fresh for each experiment by dilution of stock solutions into HNN buffer. Aliquots of stock solutions are prepared and stored as follows. 20 % (vol/vol) NP-40 in milli-Q water can be stored at room temperature, in the dark (wrap with aluminum foil) for several weeks. 200 mM Na_3VO_4 in H_2O (100×) is aliquoted and stored at -20°C. 1 M PMSF in 70% EtOH (100×) is aliquoted and stored at -20°C. Sigma protease inhibitors are aliquoted to 20ul and stored at -20°C. Aliquots are stable for at least 3 months.

C-18 elution buffer C-18 elution buffer is 50% ACN in 0.1 % formic acid.

Sample resuspension solution Sample resuspension solution is 2% ACN in 0.1% formic acid with Biognosys iRT peptides spiked in at a ratio of 1:20.

NanoLC mobile phase A NanoLC pump mobile phase A is 2% ACN in 0.1% formic acid. Mobile phase should be freshly prepared.

NanoLC mobile phase B NanoLC pump mobile phase B is 90% ACN in 0.1% formic acid. Mobile phase should be freshly prepared.

Equipment Setup

Off-line complex fractionation by size exclusion chromatography Note that SEC reproducibility can be compromised by differences in flow due to leakage or increased backpressure from clogged guard column cartridges. Monitor typical overall system backpressure and in-run backpressure to spot and solve leaks. We recommend employing two guard cartridges in line and replacing the upstream cartridge as soon as the Δp of the guard column exceeds 10 bar at 500 µl/min). To avoid damage to the 3 µm bead SEC column, avoid sudden pressure changes by adjusting flow rates only in small increments of 100 µl/min and

allowing ca. 5s for pressures to adjust. The column is equilibrated by 10 column volumes (150ml) of SEC buffer. In proteome-wide SEC fractionations, secondary interactions with the stationary phase of a subset of analytes and consequent column conditioning and washout effects cannot be avoided. Therefore, we recommend to pre-condition the column with a lysate similar to the lysate to be analyzed in SEC-SWATH-MS to ensure consistent analyte elution volume and recovery in the fractionation. Mild lysates concentrated for SEC column conditioning in 1000µg aliquots can be stored at -80°C for several months. Thaw an aliquot, spin out precipitates by 5 min of centrifugation of 16,900 ×g (4°C) and run in SEC. After column conditioning (ca. 90 min to allow full baseline equilibration of the OD signal), analyze the aqueous SEC standard sample to finish setup for analysis of the real sample(s).
Fractionations are run at 500 µl/min to minimize shear forces and with column temperature controlled at ≤ 4°C to minimize dilution-induced complex disassembly. Mobile phases are stored at room temperature and cooled by flow through the temperature-controlled autosampler module as well as heat exchanger units before entering the guard and analytical columns. To accommodate both precolumn and analytical column, we modified the housing of the column compartment with a drill. Alternatively, if no modifications are possible or if larger capacity SEC columns are to be employed, temperature can be controlled by submerging both pre-column and main column in an ice water bath.

On-line nanoLC-MS/MS Peptide samples are loaded onto a self-packed C-18 reversed phase column (75 µm ID PicoFrit emitter packed with 20 cm Magic AQ 3µm C-18 resin) at 300 nl/min and subsequently eluted by a linear gradient of 3-35% mobile phase B in mobile phase A over 120 min at 300 nl/min, with direct electrospray into the ion source and mass spectrometer. Other comparable C18 phases and nanoLC setups can be employed. The TripleTOF 5600+ mass spectrometer is operated in either data-dependent (also termed information-dependent) acquisition mode (DDA) or DIA/SWATH™2.0 acquisition mode. For detailed acquisition parameters, see Tables 1, 2 and 3. Other mass spectrometric platforms, such as e.g. the TripleTOF 6600, have been used successfully to record high quality SEC-SWATH-MS datasets.

Windows PC for file conversion For file conversion, a windows computer with a recent version of the ProteoWizard suite (≥ Version: 3.0.19228-a2fc6eda4) is required. Download and Install ProteoWizard from http://proteowizard.sourceforge.net/, as described[45].

Workstation computer for peptide-centric SWATH-MS analysis In this protocol, we employ a docker container that provides a stable solution for running peptide-centric scoring by OpenSWATH, PyProphet and TRIC on different computing systems. The workflow presented here has been tested on Linux, Windows and OSX environments. First, it is necessary to install docker (https://docs.docker.com/). Note that it might, depending on the dataset and library size, be required to extend the resources allocated to the docker software. We recommend to minimally allocate 6 CPUs, ~12000 MB memory and ~12 GB disk image size. If a task running within a docker container is suddenly "killed" without a more specific error, this usually means that not enough memory was allocated. Settings can be changed when clicking on the docker symbol in the taskbar, selecting the settings option and going to the advanced tab. To test the successful docker installation open a command line interpreter on your computer and type the following command:

```
docker run hello-world
```

Now install the OpenSWATH docker container:

```
docker pull openswath/openswath:0.1.2
docker run -u 0 -dit --name openswath -v $PWD/:/data openswath/openswath:0.1.2
```

Test if the OpenSWATH docker installation worked:

```
docker exec openswath echo hi there, openswath container is happy and alive
```

R environment and *CCprofiler* installation All data analysis in R can be performed either on a local computer or on a cluster system. To install R, download the latest release version of

R (>=3.6.0) from http://cran.r-project.org/ and install it according to the R installation and administration manual https://cran.r-project.org/doc/manuals/R-admin.html. To install the *CCprofiler* package from GitHub, you need to start the R program. For both Windows and OSX this means double-clicking on the R application icon. On UNIX-like systems you need to type 'R' in a shell prompt. Users may also want to consider using the RStudio environment (https://rstudio.com/).
Once in the R environment, run the following commands to first install and then load the devtools, data.table and *CCprofiler* packages:

```
install.packages('devtools')
library('devtools')
install.packages('data.table')
library('data.table')
install_github('CCprofiler/CCprofiler')
library('CCprofiler')
```

Procedure

Isolation of native proteome, SEC fractionation and preparation for MS analysis

Cell culture and harvest TIMING: ~7 days

1. Culture cells as applicable to the respective cell type. If using HEK293 cells, culture cells in DMEM medium supplemented with 10% FBS and 50 µg/mL penicillin/streptomycin in 15 cm cell culture dishes, incubating at 37˚C, 5% CO2. To establish a log-linearly growing cell population, split the cells twice at a ratio of 1:2 using 1× Trypsin-EDTA for 5 min at 37˚C.

2. Harvest the cells at ~80% confluency, as determined by visual inspection under the microscope. Harvest cells on ice in ice-cold PBS buffer containing 5nM EDTA using pipette flow (sufficient in the case of HEK293 cells) or a plate scraper into a 15 ml Falcon tube. Spin at 4°C, 500×g for 5min, remove supernatant using a serological pipette, and snap-freeze the cell pellet in liquid nitrogen.

PAUSE POINT: Cell pellets can be stored at -80˚C for several weeks prior to SEC-SWATH-MS analysis.

Native lysis and fractionation by size exclusion chromatography

3. Lyse cells or tissue amount sufficient to extract at least 1 mg of total protein (in the case of HEK293 cell line, 7e7 cells). Lyse cell pellets snap-frozen in step 2 by freeze-thawing into 1 ml of HNN lysis buffer. Thaw and dissolve the frozen pellet by pipetting up and down 20 times. Incubate on ice for 5 min. Other cell or tissue types may be used, whereas input amounts need to be adapted based on cell size or protein yield with a minimal pure protein amount of 600µg required as input to SEC fractionation, with concentration determined colorimetrically (e.g. using the Pierce BCA protein assay kit). This corresponds to ~2mg when protein concentration is estimated by OD280 measurements which are confounded by other molecules in the sample but used here for the sake of processing speed.

4. Fill the lysate to a volume of two milliliters with HNN lysis buffer and distribute two Ultracentrifuge tubes. Balance weight on a fine balance with HNN Lysis buffer.

5. Transfer to the pre-cooled centrifuge rotor and clarify by 15 minutes of ultracentrifugation (100,000×g, 4°C, 55,000rpm on TLA120.2 rotor).

6. Pre-cool two Amicon Ultra-4 Centrifugal Filter Units on ice. Transfer 300 µl of the cleared lysate to each Amicon device and exchange buffer to HNN buffer as follows.

CAUTION: Avoid transfer of lipids from the top layer of the supernatant by aspirating the cleared lysate from 1 cm below the liquid surface.

7. Exchange buffer to HNN buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 50 mM NaF) at a final ratio of 1:50 in three dilution and re-concentration steps to avoid large dilution steps in the interest of complex integrity. Centrifugation is performed at 3220×g, 4°C.

CRITICAL: Local precipitation occurs at and blocks the filtration membrane. It is therefore important to flush the membrane with the dilution buffer and using a 200µl pipette tip to achieve thorough rinsing of the membrane.

7.1. Centrifuge for 5' (final volume above filter ca. 200µl)
7.2. Dilute 1:5 in HNN (add 800 µl), flush membrane
7.3. Centrifuge for 10' (final vol. ~250µl)
7.4. Dilute 1:5 in HNN (add 1000 µl), flush membrane
7.5. Centrifuge for 10' (final vol. ~250µl)
7.6. Dilute 1:2 in HNN (add 250µl), flush membrane
7.7. Centrifuge 5' (vol. ~150µl), flush membrane
7.8. Centrifuge 5'
7.9. Final volume per tube: ca. 50-80µl.
7.10. Remove precipitates by centrifugation at 16,900 ×g, 4°C, for 5min, transferring the supernatant, leaving 10µl, to a pre-cooled injection vial.

8. Measure the concentration of the lysate by UV/Vis photospectrometry (Nanodrop Spectrophotometer) against a reference sample of HNN Lysis buffer in HNN buffer (1:50), approximating 1 OD280 = 1µg/µl protein concentration. The measured concentration should typically be between 20 and 30 µg/µl.

CAUTION: The concentration read by UV/Vis photospectrometry is confounded by other compounds with absorbance at 280nm. Based on colorimetric methods (BCA assay) the protein loading is ca. 3-4-fold lower than approximated by UV-Vis (Figure 2A). We suggest the fast UV-Vis reading to be sufficient to align sample loading amounts and preferable over BCA or similar quantitative assays with significant incubation times that may affect complex stability.

9. Subject 1000µg of the concentrated lysate to SEC fractionation at 500 µl/min. Ensure that the chromatographic system and column show reproducible and expectable performance in the fractionation of the protein standard mix prior to and after the analysis. Collect fractions in the expectable elution range from 10-28min at 0.19 min per fraction into a cooled 1ml 96-DeepWell plate.

10. Repeat step 9 while collecting fractions in a new 96-well plate.

11. Interrogate the UV/Vis profiles of the two SEC runs of the same lysate and if in agreement, pool the collected fractions across the two replicate injections to obtain one set of fractions.

CRITICAL: It is important to sample chromatographic fractions also of the void volume peak, even if the information of contained analyte size is reduced. This is especially important for quality control measures of the overall global proteome assembly state of the investigated cell system (observed total MS signal in assembled vs. monomeric SEC range). Additionally, the peak detection algorithms employed in downstream protein and protein complex detection benefit from complete elution profiles including shoulder regions of detectable peaks. The right boundary of the relevant protein elution range can be established empirically by SDS PAGE analysis of the late fractions (> F70). We suggest to use the elution volume of the small molecule uridine contained in the SEC standard sample. We recommend sampling until inclusive of uridine peak elution as a subset of proteins and complexes may display secondary interactions with the stationary phase and thus delayed elution in this fraction range.

12. To monitor SEC stability and to calibrate the apparent molecular weight per SEC fraction, analyze 5µl of the SEC column performance check standard after the SEC experiment.

13. Transfer an aliquot of the unfractionated sample to the collection plate. Pipette 1/40th of the volume injected for SEC (25µg by OD280) into wells H11&H12 and fill to 200µl with SEC buffer to align digest conditions with the individual SEC fractions.

CRITICAL STEP: Include an aliquot of the unfractionated sample in the proteomic analysis to ensure comparable digest conditions as for the chromatographic fractions. The data acquired from the unfractionated mild proteome is used in the PyProphet machine learning step in peptide-centric analysis, generating one scoring function applied across all chromatographic fractions to ensure aligned scoring and consistent quantification of peptides across all chromatographic fractions.

PAUSE POINT: Undigested SEC fractions can be stored at -80˚C for several weeks. Optionally, if extended storage is desired, it is recommended to denature proteins by boiling in sodium deoxy-cholate (next step) before freezing for storage.

Tryptic digest and C-18 cleanup of chromatographic fractions for MS analysis. TIMING: 4+12h

14. Denature proteins by adding sodium deoxy-cholate to 1 % v/v (20µl from 10% stock solution) and incubate 5min in a hot water-bath (95˚C).

CAUTION: Ensure that the plate is properly sealed before incubation in the water bath to avoid sample loss or contamination.

15. Let plate cool to room temperature and centrifuge at 500×g to collect liquid at the base of the plate.

16. Reduce proteins by adding TCEP to 5 mM (22 µlfrom 50 mM solution, 1:10 dilution of 500 mM stock in Ammonium bicarbonate 50mM pH 8.8). Incubate 30min at room temperature.

CRITICAL STEP: Ensure that the TCEP stock solution is titrated to pH 8.8 to avoid acidification of the samples and premature precipitation of sodium deoxy-cholate.

17. Alkylate proteins by adding iodo-acetamide (IAA) to 10 mM (24 µlfrom 100 mM stock. Incubate 20min at room temperature, in the dark.

CAUTION: Work in reduced light conditions and incubate in the dark due to IAA light sensitivity.

CRITICAL STEP: Ensure that the pH is ≥ 8.0 to avoid gel formation or partial precipitation of deoxy-cholate during the digest. Test the samples for gel formation using a pipette tip and if very high viscosity or formation of a gel are observed, adjust the pH by adding NaOH (In steps of 5 µlof 2M stock solution until the samples display low viscosity and pH 8.0 - 8.5).

18. Add 0.2 µg trypsin (Promega) per fraction (2 µl of 0.1 µg/µl stock in Trypsin buffer). Re-seal plate, shake, spin down for 1 min at 2,000×g & incubate over night at 37 °C.

19. Stop the digest and precipitate deoxycholic acid by adding TFA to 1 %, ACN to 1 % (26µl of 10% TFA / 10% ACN stock). Close and mix the plate thoroughly using a new plate seal and 10 inversions. Spin down for 1 min at 2,000×g.

20. Prepare MacroSpin plate for C-18 cleanup. Tap plate to loosen resin material and spin down for 1 min at 1000 ×g. Activate resin by adding 200 µlACN per well and centrifuging at 1,000×g for 1 min. Equilibrate the resin by 3 washes with 150ul 5 % ACN/0.1 % FA spinning at 1,000×g for 2 min. Discard washing solution from the collection plate.

21. Directly before loading the samples for C-18 cleanup, pellet the precipitated deoxycholic acid for 10 min at 3,220×g. Transfer 80 % (220 ul) of the cleared supernatant onto the equilibrated C-18 resin.

CAUTION: Ensure minimal transfer of precipitate onto the C18 resin to avoid sample contamination.

22. Load samples at 1000×g for 2 min. To maximize recovery, re-load the flow-through onto the C-18 resin a second time. Keep the flow-through for potential trouble-shooting.

23. Wash the C-18 resin by 3×200µl 5% ACN/0.1% FA, spinning at 1,000×g for 1 min each.

24. Elute the samples into a fresh collection plate with 2×150µl 50%ACN/0.1% FA.

25. Dry samples in a speed-vac equipped with a plate rotor and adequate tara plate filled with the same volume of C-18 elution buffer (45˚C, 0.2 atm, ca. 4h).

PAUSE POINT: Dried peptide samples can be stored for several weeks at -20 or -80˚C.

MS analysis: TIMING: 12h (QC) + 14 days (DIA only) OR 28 days (DIA + DDA)

26. Re-suspend dried peptide samples in 18µl 2% ACN/0.1% FA, supplemented with internal retention time calibration peptides (iRT kit, Biognosys, CH, 1:20 dilution as opposed to manufacturer's instruction of 1:10 to accommodate larger injection volumes). The spiked in iRT peptides allow the normalization of retention times across different LCMS runs and enable the streamlined generation of spectral libraries and queries of peptides from repository-scale spectral libraries in the DIA/SWATH data maps [26,31]. Re-suspend the samples by 5 min sonication in an ice-cooled water bath to avoid sample heating and evaporation.

27. Collect liquid and remove potential residual deoxy-cholate by centrifugation at 3,220×g for 5min. Transfer 16µl of the sample to MS injection vials.

CAUTION: Transfer the peptide samples pipetting at an angle and leaving ca. 2µl in order to avoid transfer of potential residual deoxy-cholate precipitate from the lowest points of the wells.

28. Before analyzing the full set of fractions, test sample set quality by analyzing 2ul of the unfractionated sample and the two fractions with highest absorbance at OD280 as monitored during SEC fractionation (In our chromatographic setup, fractions 5 and 50).

    Judge sample quality based on the following criteria:

    - no increase of chromatographic backpressure
    - TIC signal intensity in SWATH64vw mode is ≥ 2e7 (120min gradient) (Figure 2B)
    - The m/z map is well-populated with isotopic envelopes

    To acquire the full dataset, maximize sample injection volumes to target 1e8 in the highest-abundant SEC fraction (In the HEK293 case, fraction 50 and an injection volume of 4 µl).

TROUBLESHOOTING

29. If a project-specific spectral library should be generated, each fraction should be analyzed in both data-independent SWATH and data-dependent acquisition mode.

CAUTION: Datasets acquired exclusively in SWATH acquisition mode can typically be interpreted using spectral libraries from public repositories. Note that the library employed for interpretation needs to be representative for the tissue type that is being analyzed. Depending on the availability of such libraries and the research question at hand it might further be preferable to generate a project-specific spectral library by DDA acquisition of a subset of or the full sample set analyzed by SEC-SWATH-MS.

Peptide-centric SWATH-MS analysis: TIMING: 3 days

Here, we employ a docker container (see installation and initialization in Equipment Setup) that provides a stable solution for running peptide-centric scoring by OpenSWATH, PyProphet and TRIC on any computing system. Example files and a script including all processing steps are provided in our GitHub repository (https://github.com/CCprofiler/SECSWATH_PeptideCentricAnalysis.git).

30. Create a data analysis folder
    i. Open a command line interpreter
    ii. Clone and enter our analysis folder template from GitHub:
    ```
    git clone https://github.com/CCprofiler/SECSWATH_PeptideCentricAnalysis.git
    cd SECSWATH_PeptideCentricAnalysis
    ```
31. Prepare all required input data for peptide-centric analysis
    i. MS file conversion and centroiding
    On the conversion computer, use MSconvert to convert and centroid .wiff raw files into .mzML or mzXML format [45].

i. Open MSconvertGUI

ii. Under Files/browse, select the .wiff files.

iii. Under Options, leave the defaults and activate in addition 'Package in gzip'.

iv. Under 'Filters', select 'Peak Picking'.

v. Under 'Algorithm', select 'Vendor'.

vi. Under 'MS Levels', enter '1-2'.

vii. Hit 'Add'.

viii. Start the conversion (Button in the lower right).

ix. Once the conversion is finished, move the.gz file(s) to the peptide centric analysis computer and into the folder

1. `SECSWATH_PeptideCentricAnalysis/data_dia/`
2. Then, move the .gz files generated from the unfractionated sample into the subfolder
3. `SECSWATH_PeptideCentricAnalysis/data_dia/unfractionated_secinput/`

NOTE: The centroiding significantly reduces file size and processing time and is highly recommended, in particular if peptide-centric analysis is to be performed on a personal or laptop computer.

ii. Information on retention time calibration peptides (iRT spike-in or ciRT peptide set)
Example iRT and ciRT libraries are provided in the data_library folder in the cloned GitHub repository.

iii. Prepare a file specifying the SWATH window settings
An example file with SWATH window settings is provided in the data_library folder in the cloned GitHub repository (also see Table 3).

iv. Prepare a spectral library

a. Create a sample-specific spectral library according to the previously published protocol by Schubert et al. [26]

b. Download a public library such as the combined human assay library that we used for our analysis here [31]

```
wget -O data_library\spectrast2tsv.tsv
https://db.systemsbiology.net/sbeams/cgi/downloadFile.cgi?name=phl0
04_canonical_s64_osw.csv;format=tsv;tmp_file=8becf7ae782dd305c0eade
59f282bcd1;raw_download=1
```

32. Initialize the OpenSWATH docker container (see installation in Equipment Setup or follow instructions in https://github.com/CCprofiler/SECSWATH_PeptideCentricAnalysis/blob/master/SECSWATH_PeptideCentricAnalysis.sh)

```
docker attach openswath
```

33. Prepare the spectral library for OpenSWATH and PyProphet analysis

i. Convert library to .pqp file format recommended for OpenSWATH

```
TargetedFileConverter -in /data/data_library/spectrast2tsv.tsv \
-out /data/data_library/spectrast2tsv.pqp
```

ii. Generate decoys for scoring and FDR estimation in PyProphet

```
OpenSwathDecoyGenerator -in /data/data_library/spectrast2tsv.pqp \
-out /data/data_library/spectrast2tsv_td.pqp
```

34. Peptide-centric signal detection with OpenSWATH

i. Run OpenSwath on unfractionated input sample(s)

```
for file in /data/data_dia/unfractionated_secinput/*ML.gz; do \
bname=$(echo ${file##*/} | cut -f 1 -d '.'); \
OpenSwathWorkflow \
-in /data/data_dia/$bname.*ML.gz \
-tr /data/data_library/spectrast2tsv_td.pqp \
-tr_irt /data/data_library/irtkit.TraML \
-min_upper_edge_dist 1 \
-batchSize 1000 \
-out_osw /data/results/$bname.osw \
-Scoring:stop_report_after_feature 5 \
-rt_extraction_window 600 \
-mz_extraction_window 30 \
-ppm \
-threads 6 \
-use_ms1_traces \
-Scoring:Scores:use_ms1_mi \
-Scoring:Scores:use_mi_score ; done
```

ii. Run OpenSWATH on fractionated samples

```
for file in /data/data_dia/*ML.gz; do \
bname=$(echo ${file##*/} | cut -f 1 -d '.'); \
OpenSwathWorkflow \
-in /data/data_dia/$bname.*ML.gz \
-tr /data/data_library/spectrast2tsv_td.pqp \
-tr_irt /data/data_library/irtkit.TraML \
-min_upper_edge_dist 1 \
-batchSize 1000 \
-out_osw /data/results/$bname.osw \
-Scoring:stop_report_after_feature 5 \
-rt_extraction_window 600 \
-mz_extraction_window 30 \
-ppm \
-threads 6 \
-use_ms1_traces \
-Scoring:Scores:use_ms1_mi \
-Scoring:Scores:use_mi_score ; done
```

NOTE: OpenSWATH creates several warnings and errors that can be ignored when analyzing SEC-SWATH-MS datasets, including:

- Warning "windows were sparce" and/or "empty chromatogram": Sparsity of certain windows is expected for some fractions, especially in the beginning and end of the SEC.
- Error "Transition does not have a corresponding chromatogram"

35. Peptide-centric scoring with PyProphet

a. Train Model: pyProphet analysis of unfractionated sample

```
pyprophet score
--threads 6
--in=/data/results/unfractionated_secinput/unfractionated_secinput.osw \
--out=/data/results/unfractionated_secinput/model.osw
--level=ms1ms2
```

b. Apply global model to score peak groups in all runs evenly

i. Scoring and plotting

```
for file in /data/results/*.osw; do \
bname=$(echo ${file##*/} | cut -f 1 -d '.'); \
pyprophet score --in=/data/results/$bname.osw \
--apply_weights=/data/results/unfractionated_secinput/model.osw \
--level=ms1ms2; done
```

ii. Exporting of output files

```
for file in /data/results/*.osw; do \
bname=$(echo ${file##*/} | cut -f 1 -d '.'); \
pyprophet export --in=/data/results/$bname.osw \
--out=/data/results/$bname.tsv \
--max_rs_peakgroup_qvalue=0.1 \
--no-transition_quantification \
--format=legacy_merged; done
```

NOTE: We advise to manually check if .tsv output files are actually written for all runs.

iii. Plotting of all score distributions

```
for file in /data/results/*.osw; do \
bname=$(echo ${file##*/} | cut -f 1 -d '.'); \
pyprophet export --in=/data/results/$bname.osw \
--format=score_plots; done
```

36. TRIC based feature alignment across all SEC fractions

```
feature_alignment.py \
--in /data/results/*.tsv \
--out /data/results/feature_alignment.tsv \
--out_matrix /data/results/feature_alignment_matrix.tsv \
--method LocalMST \
--realign_method lowess \
--max_rt_diff 60 \
--mst:useRTCorrection True \
--mst:Stdev_multiplier 3.0 \
--target_fdr -1 \
--fdr_cutoff 0.05 \
--max_fdr_quality 0.1 \
--alignment_score 0.05
```

SEC-SWATH-MS data processing and complex-centric analysis in *CCprofiler*: TIMING 2 days

CRITICAL STEP: Part 3 of the PROCEDURE describes how to use the open-source *CCprofiler* R-package to extract information about the global proteome assembly state and specific protein complexes from co-fractionation MS experiments, here generated by SEC-SWATH-MS. The analysis includes: data import and pre-processing (Steps 34-37), automated parameter selection (Step 38), protein-centric analysis (Step 39) and complex-centric analysis (Step 40). All *CCprofiler* analysis steps are also provided as a supplementary R-script that performs the presented analysis based on the exemplary HEK293 SEC-SWATH-MS dataset. The R-script can easily be adapted to other datasets by changing the input files (Step 34-35). All exemplary data and the script are available on GitHub: https://github.com/CCprofiler/SECSWATH_ComplexCentricAnalysis (also see the Supplementary Manual).

To set up your work environment you can clone the GitHub repository by:

```
git clone https://github.com/CCprofiler/SECSWATH_ComplexCentricAnalysis.git
cd SECSWATH_ComplexCentricAnalysis
```

NOTE: Due to parallelization of some of the *CCprofiler* processing steps and involved random number generation that is beyond our control, the results of the workflow are subject to minor variation despite setting a seed value. If fully reproducible results are desired, only a single processing core should be selected. This is however connected to much longer processing times.

PAUSE POINT: The following computational analysis can essentially be paused at any point when a certain function is completed. Before closing R it is important to save the environment in order to resume the analysis at a later stage. For this, use the following commands:

```
save.image(file='CCprofiler_analysis.RData')
```

To resume your analysis, you can load the previous status of your R environment with the following command:

```
load(file='CCprofiler_analysis.RData')
```

34. **Prepare data for *CCprofiler* import**
    Prepare all necessary data that needs to be loaded into R for the *CCprofiler* analysis. For convenience we recommend saving all input data in the same directory where you want to perform the analysis. All data necessary and used for this protocol are provided in the GitHub repository and will be available in the SECSWATH_ComplexCentricAnalysis folder after you cloned it (see above).
    i. Prepare quantitative peptide-level data
        A. Quantitative peptide matrix generated by OpenSWATH (as described in Part 2)
            i. The output table from TRIC can directly be imported into *CCprofiler* (see ‘feature_alignment.tsv’ or ‘quantData_OpenSWATH.rds’ (already in R data format))
        B. Quantitative peptide matrix generated by any software tool
            i. Remove decoys
            CAUTION: Decoys might be valuable for certain processing steps downstream (e.g. selecting a sibling peptide correlation based FDR cutoff). We have specifically tested the propagation of decoys for datasets processed by an OpenSWATH-based workflow. If other data processing tools have been used, the decoys should be treated with caution. To be on the conservative side, we would generally

recommend removing the decoys.

ii. Remove non-proteotypic peptides

iii. Bring data in either long or wide format:

a. Required column names for long format: protein_id, peptide_id, filename and intensity (see ‘examplePCPdataLong.tsv’)

b. Required column names for wide format: protein_id, peptide_id, <filename1>, <filename2>, …, <filenameX>
(see ‘examplePCPdataWide.tsv’)

ii. In addition to the quantitative peptide matrices, *CCprofiler* requires a fraction annotation table that maps each filename to a given chromatographic fraction number. The required column names are: filename and fraction_number (see ‘exampleFractionAnnotation.tsv‘).
CAUTION: The filenames used in the fraction annotation table need to match the filenames in the quantitative matrix exactly. Further, the fraction_number entries need to start with 1 and continuously increase in integer steps of 1 until the last sampled fraction.

iii. For native complex separation via SEC, a molecular weight (MW) calibration table can be generated by measuring the apex fractions of an external standard set of reference proteins fractionated on the same SEC setup. By providing such a MW calibration table, *CCprofiler* can establish a transformation function based on the log-linear relationship between elution fractions and apparent MWs inherent to SEC, thus enabling the annotation of all sampled fractions with an apparent MW. The required column names in the calibration table are: std_weights_kDa and std_elu_fractions (see ‘exampleCalibrationTable.tsv’).

iv. *CCprofiler* can further annotate protein traces with additional information provided in a trace annotation table, e.g. adding the gene names or monomeric MW from UniProt (https://www.uniprot.org/) (see ‘exampleTraceAnnotation.tsv’). Adding information on monomeric MWs of the analyzed proteins is critical for the assignment of proteins to monomeric or complex-assembled state from SEC datasets with calibrated apparent MW and is required for the assessment of global proteome assembly states.
CAUTION: The protein_id column in the quantitative matrix needs to match one of the column entries in the annotation table. Typically, the common entry are the UniProt identifiers.

v. Finally, a necessary component for downstream detection of protein complexes by complex-centric analysis (Step 40), is the selection of prior protein connectivity information which can be provided either in the form of defined protein complexes, e.g. as annotated in CORUM [41,46], or binary interaction networks generated by various approaches, as for example the BioPlex [1,2] or StringDB [42,43] networks.

A. Defined complex hypotheses
A table with defined complexes should contain the following columns: complex_id, complex_name and protein_id (see ‘corumComplexHypothesesRedundant.csv’).

B. Binary protein-protein interaction network
The format for a binary interaction network is a table with two columns: a and b. Both columns contain protein identifiers and each row represents a binary connection (an ‘edge’) in the interaction network (see ‘BioPlexPPIs.tsv’).

CAUTION: The protein_id / a & b entries need to correspond to the protein_id in the quantitative matrix, e.g. UniProt identifiers.

35. Load input tables into R and inspect

i. Load libraries in R

```
library(data.table)
library('CCprofiler')
```

ii. Set working directory to the location where all files are stored.

```
setwd("SECSWATH_ComplexCentricAnalysis")
```

iii. Load and inspect the quantitative peptide matrix

i. Quantitative peptide-level data generated by OpenSWATH (as described in Part 2)

```
quantData_OpenSWATH <- readRDS("quantData_OpenSWATH.rds")
```

ii. Quantitative peptide matrix generated by any software tool

a. Long format

```
quantData_long <-
fread("examplePCPdataLong.tsv")
head(quantData_long)
```

b. Wide format

```
quantData_wide <-
fread("examplePCPdataWide.tsv")
head(quantData_wide[,1:5])
```

iv. Load and inspect fraction annotation table

```
fractionAnnotation <- fread("exampleFractionAnnotation.tsv")
head(fractionAnnotation)
```

v. Load and inspect calibration table

```
calibrationTable <- fread("exampleCalibrationTable.tsv")
calibrationTable
```

vi. Load and inspect trace annotation table

```
uniprotAnnotation <- fread("exampleTraceAnnotation.tsv")
head(uniprotAnnotation)
```

vii. Load and inspect protein connectivity information

i. Defined complex hypotheses from the Corum database

```
corumComplexes <- fread("corumComplexHypothesesRedundant.csv")
head(corumComplexes)
```

ii. Binary protein-protein interaction network from BioPlex (v1.0 [1], http://bioplex.hms.harvard.edu)

```
BioPlexPPIs <- fread("BioPlexPPIs.tsv")
head(BioPlexPPIs)
```

36. **Import peptide level data into *CCprofiler* traces format and annotate**

i. **Import quantitative peptide matrix as traces object**

The traces object is the main data class used in the *CCprofiler* package. It stores the quantitative profiles ('traces') of peptide or protein intensities across the analyzed chromatographic fractions. Additionally, a traces object can store specific information about each of the peptides, proteins and chromatographic fractions. As the analysis proceeds more information will be added to the traces object.

i. Quantitative peptide level data generated by OpenSWATH

```
pepTraces <- importFromOpenSWATH(data = quantData_OpenSWATH,
annotation_table = fractionAnnotation,
verbose = FALSE)
```

ii. Quantitative peptide matrix generated by any software tool

NOTE: CCprofiler will automatically detect if peptide tables are in long or wide format.

a) Long format

```
pepTraces_exampleSubset_long <
importPCPdata(input_data = quantData_long,
fraction_annotation = fractionAnnotation,
rm_decoys = FALSE)
```

b) Wide format

```
pepTraces_exampleSubset_wide <- importPCPdata
(input_data = quantData_wide,
fraction_annotation = fractionAnnotation,
rm_decoys = FALSE )
```

ii. **Perform molecular weight calibration and annotation**

i. Perform molecular weight calibration based on a provided calibration_table (Figure 3A):

```
calibration = calibrateMW(calibration_table =
calibrationTable,
PDF = plotPDF)
```

ii. Annotate traces with the apparent molecular weight associated with each SEC fraction as extrapolated from the standard protein molecular weights and associated elution fraction numbers:

```
pepTraces <- annotateMolecularWeight(
traces = pepTraces,
calibration = calibration)
```

CAUTION: Apparent molecular weight calibration is of limited accuracy as, inherent to the analytical procedure wherein analyte shape and propensity for unintended secondary interaction with the stationary phase affect elution volumes/fraction number and inferred apparent molecular weight. Predictions, especially those outside the range of standard protein elution, should be interpreted with caution.

iii. **Annotate traces with information from UniProt**

```
pepTraces <- annotateTraces(traces = pepTraces,
                            trace_annotation = uniprotAnnotation,
                            traces_id_column = "protein_id",
                            trace_annotation_id_column = "Entry")
```

37. **Pre-process traces object to increase data quality**

i. **Optional: Detect and impute missing values**

In most proteomics pipelines, zero intensity values indicate either that the signal is missing at random (no detection due to technical reasons such as interferences from other peptides) or missing not at random (no detection due to cellular concentrations below the detection limit). We suggest that a zero value is likely missing at random in case a quantitative (non-zero) signal has been detected in both the previous and following fraction. The detected missing at random values are subsequently imputed by a spline fit across the fractionation dimension.

1. Convert zeros in missing at random value locations to NA:

```
pepTracesMV <- findMissingValues(traces = pepTraces,
bound_left = 2,
bound_right = 2,
consider_borders = FALSE)
```

2. Impute NA values by fitting a spline:

```
pepTracesImp <- imputeMissingVals(
traces = pepTracesMV,
method = "spline")
```

3. Plot imputation summary:

```
plotImputationSummary(
traces = pepTracesMV,
tracesImp = pepTracesImp,
max_n_traces = 5,
PDF = plotPDF)
```

NOTE: In the original complex-centric study of the HEK293 proteome [15] no missing values were imputed. Generally, quantitative matrices from SWATH-MS, particularly with TRIC alignment, display only few missing values and imputation thus has little influence in such datasets. However, imputation improves overall workflow robustness and flexibility for different input data types. For example, loss of data from an entire SEC fraction due to failed MS acquisition can robustly be compensated by imputation rather than re-analysis of the fraction or repeat of the entire experiment. Further, missing value imputation should improve the interpretability of datasets affected by more missing values, e.g. when acquired via classical data-dependent mass spectrometry.

ii. **Filter peptides by consecutive peptide detection**

Peptides that have never been detected in more than N consecutive fractions, here N=2, are removed from the traces object. This effectively removes false positive peptide detections from the dataset.

```
pepTracesConsIds <- filterConsecutiveIdStretches(
traces = pepTracesImp,
min_stretch_length = 3,
remove_empty = TRUE)
```

iii. **Select high-quality proteins based on their average sibling peptide correlation**

i. **Calculate the average sibling peptide correlation (SPC) for each peptide**

For each peptide, the average pairwise correlation with the quantitative traces of its sibling peptides, i.e. peptides derived from the same protein, is calculated (Figure 3B).

```
pepTracesSibPepCorr <- calculateSibPepCorr(
traces = pepTracesConsIds,
PDF = plotPDF)
```

ii. **Filter by SPC**

Peptides below a minimum average SPC cutoff are removed. The rational is that outlier peptides as well as proteins with very heterogeneous quantitative peptide traces are excluded from further analysis. The filtering cutoff can either be automatically determined by a target-decoy based FDR estimation approach (a), or a fixed cutoff can be applied (b):

a. SPC based FDR cutoff (Figure 3C)

A conservative FFT can be estimated from the unfractionated SEC input sample that was also used to train the PyProphet model for peptide-centric analysis. This is conservative, because we expect to see cumulatively more proteins in the SEC fractions than in the single unfractionated input sample. The estimated pi0 ~ FFT is reported in the protein-level pdf report. For this dataset the FFT was estimated to be 0.491.

```
estimatedFFT <- 0.491
```

Filter by FDR cutoff using the estimated FFT:

```
pepTraces_filtered_FDR <- filterBySibPepCorr(
traces = pepTracesSibPepCorr,
fdr_cutoff = 0.01,
FFT = estimatedFFT,
rm_decoys = TRUE,
PDF = plotPDF)
```

CAUTION: This option is only valid if you have continuously kept decoys in your analysis. The most conservative strategy is to then apply a FFT of 1. However, if you have a FFT estimation available this will significantly boost your sensitivity and result in a higher number of remaining proteins for the downstream analysis. We have specifically tested this option for datasets processed by an OpenSWATH-based workflow. If other data processing tools have been used, the decoy based FDR estimation on SEC level should be treated with caution.

b. Absolut sibling peptide correlation cutoff

```
pepTraces_filtered_absoluteCutoff <-
filterBySibPepCorr(
traces = pepTracesSibPepCorr,
fdr_cutoff = NULL,
absolute_spcCutoff = 0.25,
rm_decoys = TRUE,
PDF = plotPDF)
```

iv. Inspect resulting peptide-level traces object

i. Summary statistics

```
summary(pepTraces_filtered_FDR)
```

ii. Plot some example traces, here the exemplary visualization of the Proteasome subunit alpha type-1 (UniProt ID = P25786, Figure 3E)

```
test_protein <- c("P25786")
test_peptide_traces <- subset(
traces = pepTraces_filtered_FDR,
trace_subset_ids = test_protein,
trace_subset_type = "protein_id")
plot(test_peptide_traces,
PDF = plotPDF,
name = paste0("pepTraces_",test_protein))
```

v. Protein quantification

i. Perform protein quantification by selecting the top N, here N=2, peptides based on their global intensity across all fractions.

```
protTraces <- proteinQuantification(pepTraces_filtered_FDR,
topN = 2,
keep_less = FALSE)
```

ii. Inspect summary statistics of the resulting protein traces

```
summary(protTraces)
```

CRITICAL STEP: Compare the number of remaining proteins to the number of proteins on the peptide level traces. If the number of proteins is dramatically reduced during the protein quantification step, many proteins might have been detected by a single peptide only. Careful consideration is necessary to decide whether you want to trust such single peptide hits and include them in your downstream analysis by reducing the quantification criteria.

iii. Visualize and inspect example protein traces

Exemplary visualization of the Proteasome subunit alpha type-1 (UniProt ID = P25786)

```
test_protein_traces <- subset(
traces = protTraces,
trace_subset_ids = test_protein,
trace_subset_type = "protein_id")
plot(test_protein_traces,
colour_by = "Entry_name",
PDF = plotPDF,
name = paste0("protTraces_",test_protein))
```

vi. **Overall workflow QC to evaluate the global proteome assembly state**

The protein-level profiles can then be used to estimate the overall complex assembly state observed in the sample as a quality control to ensure the successful extraction and profiling of largely intact complexes. Here, we evaluate the total MS signal in assembled vs. monomeric range (Figure 3D).

```
summarizeMassDistribution(protTraces,
PDF = plotPDF)
```

38. **Automatically identify optimal processing parameters based on a protein-level parameter grid search**

A grid search can be performed to determine an optimal set of parameters for the protein- and/or complex-centric proteome profiling workflow. This optimal parameter set depends mostly on the co-fractionation characteristics and MS setup.

i. **Randomly select a subset of proteins for the grid search**

The selected subset of proteins should be representative of the proteome, thereby providing a trade-off between coverage and computational run-time. From our experience, selecting < 100 proteins suffers in regard to robustness, while >500 proteins will require a lot of processing time. We therefore propose a random selection of ~500 proteins.

```
all_proteins <-
unique(pepTraces_filtered_absoluteCutoff$trace_annotation$protein_id)
testProtein_idx <- sample(1:length(all_proteins), 500)
testProteins = all_proteins[testProtein_idx]
peptideTracesSubset = subset(
traces = pepTraces_filtered_FDR,
trace_subset_ids = testProteins,
trace_subset_type = "protein_id")
```

ii. **Perform parameter grid search**

The grid search performs a peptide co-elution peak group finding for a selected combination of parameters with the goal to determine a good parameter set for the following analyses. Please note that the selection of suitable parameters is for the grid seach is critical

```
gridFeatures <- performProteinGridSearch(
traces = peptideTracesSubset,
corrs = c(0.9,0.95),
windows = c(8,10),
smoothing = c(7,9),
rt_heights = c(1,3),
n_cores = 3)
```

CRITICAL The selection of parameters for the grid search is critical. Guidelines for the selection of reasonable parameters are discussed in Box 2.

iii. **Score protein features across all grid search parameters and select the best parameter set**

```
gridFeatures_scored <- lapply(gridFeatures,
calculateCoelutionScore)
gridFeatures_qvalues <- lapply(gridFeatures_scored,
calculateQvalue,
plot = FALSE)
gridFeatures_stats <- qvaluePositivesPlotGrid(
featuresGrid = gridFeatures_qvalues,
colour_parameter = "corr",
PDF = plotPDF)
bestParameters <- getBestQvalueParameters(
stats = gridFeatures_stats,
FDR_cutoff = 0.05)
bestParameters
write.table(bestParameters,
"bestParameters.tsv",
sep = "\t",
quote = FALSE,
row.names = FALSE)
```

CRITICAL Inspect the pseudo ROC curves generated by the grid search (Figure 4A). Optimal parameters are at the upper left corner of the observed distribution. Parameters that are consistently in the upper left corner are especially important.

39. **Perform protein-centric analysis**

Protein-centric analysis detects peptide co-elution peak groups along the chromatographic dimension. Each detected peak (‘protein feature’) represents the protein in a specific assembly state, i.e. monomeric or bound to different protein complexes.

i. **Perform protein feature finding**

```
proteinFeatures <- findProteinFeatures(
traces = pepTraces_filtered_FDR,
corr_cutoff = bestParameters$corr,
window_size = bestParameters$window,
rt_height = bestParameters$rt_height,
smoothing_length = bestParameters$smoothing_length,
collapse_method = "apex_only",
perturb_cutoff = "5%",
parallelized = TRUE,
useRandomDecoyModel = TRUE)
```

ii. **Score detected protein features and estimate FDR**

```
proteinFeatures_scored <- scoreFeatures(
features = proteinFeatures,
FDR = 0.05,
PDF = plotPDF)
write.table(proteinFeatures_scored,
"proteinFeatures_scored.tsv",
sep = "\t",
quote = FALSE,
row.names = FALSE)
```

CRITICAL STEP: Inspect the p-value density histogram (Figure 4B/C). There should be a high peak close to zero and a uniform distribution across all other p-values.

TROUBLESHOOTING

iii. **Inspect summary statistics on resulting protein features**

The resulting figures provide information about the number of unique assembly states detected for all the proteins as well as about the number of proteins with at least one assembled protein signal (MW ≥ 2x monomeric MW in SEC) (Figure 4D).

```
summarizeFeatures(feature_table = proteinFeatures_scored,
PDF = plotPDF,name = "proteinFeatures_summary")
```

iv. **Visualize and inspect protein features (Figure 4E)**

```
plotFeatures(feature_table = proteinFeatures_scored,
traces = pepTraces_filtered_FDR,
calibration = calibration,
feature_id = test_protein,
annotation_label = "Entry_name",
onlyBest = FALSE,
peak_area = TRUE,
monomer_MW = TRUE,
PDF = plotPDF,
name = paste0("protFeatures_",test_protein))
```

Plot all detected proteins

```
allDetectedProteins <- unique(proteinFeatures_scored$protein_id)
pdf("allDetectedProteins.pdf", height = 6, width = 8)
for (protein in allDetectedProteins) {
plotFeatures(feature_table = proteinFeatures_scored,
traces = pepTraces_filtered_FDR,
calibration = calibration,
feature_id = protein,
annotation_label = "Entry_name",
onlyBest = FALSE,
peak_area = TRUE,
monomer_MW = TRUE,
PDF = FALSE)
}
dev.off()
```

CRITICAL STEP: Inspect some detected protein features and evaluate if the detected peak groups correspond to what you would have also selected as

peak groups during manual inspection.
TROUBLESHOOTING

40. **Complex-centric analysis**

Complex feature finding represents the central step of complex-centric analysis using *CCprofiler*. Based on prior protein interaction data and quantitative fractionation profiles, *CCprofiler* detects groups or subgroups of locally co-eluting proteins, indicating the presence of protein-protein complexes in the biological sample. Target complex queries are supplemented with decoy complex queries to support error control of the reported results. The result is a table summarizing the presence and composition of protein-protein complexes in the biological sample analyzed.

i. **Prepare target complex queries**

There are two options for protein complex target generation in *CCprofiler*: (a) use defined protein complex models for direct use as queries (2 or more subunits, e.g. from CORUM) or (b) use a protein-protein interaction network from which target complex queries can be extracted.

a) **Inspect the coverage of pre-defined protein complex queries from the previously loaded CORUM database (Figure 5A)**

```
plotSummarizedMScoverage(hypotheses = corumComplexes,
protTraces = protTraces,
PDF = plotPDF,
name_suffix = "CORUM")
```

b) **Generate and inspect protein complex queries from binary PPI networks, here based on BioPlex**

Decoy complex queries are generated based on the target complex query set and its underlying network structure. The minimum distance specifies the minimal number of edges between any two proteins within any generated decoy complex query. It is important that the interaction network based on the targets is large enough to generate a random decoy set that does not overlap with the target complex queries. We recommend complex query sets of at least 1000 targets for the decoy based approach.

i. Calculate pairwise distances between any two proteins in the interaction network

```
pathLengthBioPlexPPIs <-
calculatePathlength(BioPlexPPIs)
```

ii. Generate protein complex targets by grouping proteins based on a user-defined distance cutoff. Here we consider only direct neighbours of each protein.

```
networkTargetsBioPlexPPIs <-
generateComplexTargets(dist_info =
pathLengthBioPlexPPIs,
max_distance = 1,
redundancy_cutoff = 0)
```

iii. Inspect newly generated protein complex queries

```
head(networkTargetsBioPlexPPIs)
plotSummarizedMScoverage(
hypotheses = networkTargetsBioPlexPPIs,
protTraces = protTraces,
PDF = plotPDF,
name_suffix = "BioPlex")
```

CRITICAL:

- It is essential that the chosen protein complex queries match the experimental dataset. Therefore, inspect the protein and complex coverage pie charts (Figure 5A). We recommend that at least half of the proteins and protein complexes represented in the complex query set should be (partially) detected in the experiment.
- One critical question during complex query generation is how to handle redundancies, i.e. protein complex queries that partially or fully overlap. Due to the complex-centric scoring functions in *CCprofiler*, we recommend to also keep protein complex subsets

in the target queries. Instead of merging / removing overlapping queries at this stage we recommend to collapse detected complex signals at Step 40vi.

- If you are especially interested in some protein complexes that are not present in any available database, you can manually append these complexes to a generated target query list. It is important to keep in mind that the target query list should always contain at least around 1000 complexes in order to ensure robust decoy based FDR estimation and sensitive detection rates. If less complex queries are selected, feature finding can still be performed, but decoy generation and FDR estimation are not applicable.

TROUBLESHOOTING

ii. **Prepare decoy complex queries**

```
binaryCorumComplexes <- generateBinaryNetwork(corumComplexes)
pathLengthCorumComplexes <-
calculatePathlength(binaryCorumComplexes)
corumComplexesPlusDecoys <- generateComplexDecoys(
target_hypotheses = corumComplexes,
dist_info = pathLengthCorumComplexes,
min_distance = 2,
append = TRUE)
```

TROUBLESHOOTING

CRITICAL: Decoy complex queries are generated based on the target complex query set and its underlying network structure. The minimum distance specifies the minimal number of edges between any two proteins within any generated decoy complex query. It is important that the interaction network based on the targets is large enough to generate a random decoy set that does not overlap with the target complex queries. We recommend complex query sets of at least 1000 targets for the decoy based approach.

iii. **Perform complex feature finding**

Protein complex features are determined similar to the protein features described above. First, a sliding window strategy is applied, where all proteins of a protein complex hypothesis are tested for local profile correlation. If a subset of the proteins within a protein complex hypothesis correlate better then the specified cutoff, a protein complex feature is initiated, followed by peak detection within the regions of high correlation.

```
complexFeatures <- findComplexFeatures(
traces = protTraces,
complex_hypothesis = corumComplexesPlusDecoys,
corr_cutoff = bestParameters$corr,
window_size = bestParameters$window,
rt_height = bestParameters$rt_height,
smoothing_length = bestParameters$smoothing_length,
collapse_method = "apex_network",
perturb_cutoff = "5%",
parallelized = TRUE,
n_cores = 3)
```

CRITICAL: If no parameter selection was performed on the protein-centric level you can also do a complex level grid search [15].

i. **Filter complex features according to their apparent molecular weight, removing protein complex features that elute at an apparent molecular weight lower than any of the monomeric molecular weights of its subunits.**

```
complexFeaturesFilteredMW <- filterFeatures(
feature_table = complexFeatures,
min_monomer_distance_factor = 2)
```

ii. **Select only the best complex feature, i.e. the complex signal with most subunits and highest correlation. This step is necessary prior to the statistical scoring, because individual elution peaks are not independent.**

```
complexFeaturesBest <- getBestFeatures(
feature_table = complexFeaturesFilteredMW)
complexFeaturesBest_scored <- scoreFeatures(
features = complexFeaturesBest,
FDR = 0.05,
PDF = plotPDF,
name = "complex_qvalueStats")
summarizeFeatures(complexFeaturesBest_scored,
PDF = plotPDF,
name = "complexFeaturesBest_feature_summary")
```

CRITICAL STEP: Inspect the p-value density histogram (Figure 5B/C). There should be a high peak close to zero and a uniform distribution across all other p-values.

TROUBLESHOOTING

iii. **Append secondary features based on a user defined local subunit correlation cutoff, here 0.5.**

```
complexFeaturesAll <- appendSecondaryComplexFeatures(
scoredPrimaryFeatures = complexFeaturesBest_scored,
allFeatures = complexFeaturesFilteredMW,
peakCorr_cutoff = 0.5)
write.table(complexFeaturesAll,
"complexFeaturesAll.tsv",
sep = "\t",
quote = FALSE,
row.names = FALSE)
```

iv. **Inspect summary statistics on resulting protein features (Figure 5D)**

```
summarizeFeatures(complexFeaturesAll,
PDF = plotPDF,
name = "complexFeaturesAll_feature_summary")
plotSummarizedComplexes(
complexFeatures = complexFeaturesAll,
hypotheses = corumComplexes,
protTraces = protTraces,
PDF = plotPDF)
```

v. **Visualize and inspect detected complex features (Figure 5E)**

```
testComplex <- "181"
plotFeatures(feature_table = complexFeaturesAll,
traces = protTraces,
calibration = calibration,
feature_id = testComplex,
annotation_label = "Entry_name",
onlyBest = FALSE,
peak_area = TRUE,
monomer_MW = TRUE,
PDF = plotPDF,
name = paste("complexFeatures_",testComplex))
```

Plot all detected complexes

```
allDetectedComplexes <- unique(complexFeaturesAll$complex_id)
pdf("allDetectedComplexes.pdf", height = 6, width = 8)
for (complex in allDetectedComplexes) {
plotFeatures(feature_table = complexFeaturesAll,
traces = protTraces,
calibration = calibration,
feature_id = complex,
annotation_label = "Entry_name",
onlyBest = FALSE,
peak_area = TRUE,
monomer_MW = TRUE,
PDF = FALSE)
}
dev.off()
```

CRITICAL STEP: Inspect some detected complex features and evaluate if the detected peak groups correspond to what you would have also selected as peak groups during manual inspection.

TROUBLESHOOTING

vi. **Collapse overlapping and redundant co-elution evidence to delineate complexes and complex families with defined co-elution of subunits in SEC**

```
complexFeaturesUnique <- getUniqueFeatureGroups(
feature_table = complexFeaturesBest_scored,
rt_height = 0,
distance_cutoff = 1.25)
complexFeaturesCollapsed <- callapseByUniqueFeatureGroups(
feature_table = complexFeaturesUnique,
rm_decoys = TRUE)
write.table(complexFeaturesCollapsed,
"complexFeaturesCollapsed.tsv",
sep = "\t",
quote = FALSE,
row.names = FALSE)
```

CRITICAL STEP: To retrieve unique , non-redundant protein complex signals, the reported complex signals need to be collapsed based on a strategy that considers (i) subunit composition and (ii) resolution in the chromatographic dimension.

vii. Visualize and inspect all collapsed complex features

```
allCollapsedComplexes <-
unique(complexFeaturesCollapsed$complex_id)
pdf("allCollapsedComplexes.pdf", height = 6, width = 8)
for (complex in allCollapsedComplexes) {
plotFeatures(feature_table = complexFeaturesCollapsed,
traces = protTraces,
calibration = calibration,
feature_id = complex,
annotation_label = "Entry_name",
onlyBest = FALSE,
peak_area = TRUE,
monomer_MW = TRUE,
PDF = FALSE)
}
dev.off()
```

Troubleshooting

Step	problem	possible reason	solution
28	Missing or poor quality MS data of an isolated chromatographic fraction.	Problems in sample workup, NanoLC failure, data corruption	• The (re-) analysis of the fraction can be skipped and intensities extrapolated from adjacent fractions that were analyzed successfully. This can be achieved by using the function imputeMissingValues of *CCprofiler* (see Step 37i). It is important that missing runs are still included in the traces import to ensure consecutive fraction numbers prior to imputation.
39ii	p-value histogram for protein signals does not show a uniform distribution	• Too few queries were tested • inappropriate parameters for peak detection	• Check if the number of proteins with ≥ 2 peptides is ≥ 500 • Refer to Box 2 for guidelines on parameter selection and parameter screening (grid-search)
39iv & 40v	Detected protein / protein complex signal apex and boundaries do not look reasonable	• Parameters for the feature finding were not selected appropriately	• Check automated parameter grid search again • Manually evaluate if selected parameters are reasonable (also see Box 2) • Maybe try a parameter selection on protein complex level
40i & ii	Error in target or decoy complex query generation	• Selected binary interaction network is too small or not connected	• Use matching interaction network for your sample
40ii	p-value histogram for protein complex signals does not show a uniform distribution	• Too few queries were tested • Too few queries are observable in your dataset • Error in protein complex decoy generation	• Increase the number of protein complex queries (> 1000 queries) • Manually plot the protein profiles of a few standard complexes (e.g. proteasome, CTT complex) to test if any good signals can be observed • Check if the selected parameters are sensible • Check if decoys are present

Anticipated results

We used the presented protocol to study the proteome assembly state of a population of exponentially growing HEK293 cells [15]. In this study, we observed 5124 proteins at 1% protein FDR (Figure 3C), using the SEC-informed filtering approaches as presented above. We could show that 64% of the proteins are present in at least one assembled state according to the molecular weight distribution along the SEC and that 27% of proteins distribute into multiple distinct assembly states, as evidenced by unique elution peaks (Figure 4D). Using CORUM as prior information for complex-centric analysis, we could observe evidence for 574 protein complexes at 5% FDR (Figure 5E), boiling down to 195 unique protein complex signals after feature collapsing. We further demonstrated improved coverage of observable protein complexes by combining prior connectivity information from multiple sources. Furthermore, we demonstrated the sub-complex resolution of protein complex information retrievable by complex-centric proteome profiling. Specifically, the study identified a novel sub-complex of the COP9 signalosome complex (holo-CSN) which, due to the absence of the catalytically active subunit CSN5 and presence of subunits involved in substrate recruitment, may be able to attenuate holo-CSN de-neddylation activities by competitive binding to and sequestration of Cullin-Ring ligase substrate complexes. Second, complex-centric profiling revealed the specific composition and relative abundance of an unexpected, late stage intermediate of 20S proteasome assembly, the composition of which suggests an alternative sequence of subunit assembly when compared to the canonical model of 20S assembly [47,48]. Further, the quantitative distribution of proteins across these and other instances of SEC-resolvable protein complexes was assessed.

It can generally be anticipated that complex-centric proteome profiling of human cell lines results in the detection of at least 50,000 proteotypic peptides and 4000 uniquely detectable proteins (at 1% protein FDR). More than 50% of the protein mass (estimated from total MS signal) are expected to be detectable in likely complex-assembled state (appearing with ≥ 2x the monomeric MW in SEC). Protein-centric detection of co-elution signals from peptide level chromatograms should yield ~1-6 high-quality protein elution peaks for at least 80% of the detected proteins (q-value cutoff = 0.05, equivalent to 5% FDR). In our experience with human cell lines, ca. 25 % of the proteins distribute into multiple distinct assembly states, as evidenced by multiple resolved protein elution peaks along the fractionation dimension. Two thirds of the proteins are observed at least once eluting in assembled state (peak apex at a MW ≥ 2x the monomeric MW in SEC). Using the CORUM protein complex database [41] as prior connectivity information for complex-centric analysis in *CCprofiler*, human SEC-SWATH-MS data analysis can yield evidence for approximately 400 protein complexes at 5% FDR, boiling down to evidence for ~200 unique protein complex signals after feature collapsing. Explanations of the result tables exported by *CCprofiler* are provided in Table 5 (protein-centric analysis results), Table 6 (complex-centric analysis results) and Table 7 (results of complex feature collapsing).

Contributions

I.B., M.H. and R.A. wrote the manuscript with input from all authors; I.B. and M.H. developed the presented workflow, implemented the analysis scripts and performed all analyses; M.H. developed and optimized the experimental protocol for SEC-SWATH-MS; GR and MH optimized the peptide-centric analysis for SEC-SWATH-MS applications; I.B., M.H., M.F., G.R., R.H. and A.B.E. developed the *CCprofiler* software, R.A., MG, BCC and MH conceptualized the primary study; B.C.C., M.G. and R.A. supervised the study.

Acknowledgments

The project was supported by the SystemsX.ch projects PhosphoNetX PPM and project TbX to R.A., and the European Research Council (ERC-20140AdG 670821 to R.A.). M.H. was supported by a grant from Institut Mérieux. I.B. was supported by the Swiss National Science Foundation (grant no. 31003A_166435). B.C.C. was supported by a Swiss National Science Foundation Ambizione grant (PZ00P3_161435). A.B.E. was supported by the National Institutes of Health project Omics4TB Disease Progression (U19 AI106761).

Competing interests

The authors declare that they have no conflict of interest.

Data and Software availability

The mass spectrometry data for the HEK293 SEC-SWATH-MS experiment[15] is available at ProteomeXchange Consortium PXD007038 (http://proteomecentral.proteomexchange.org)

A detailed protocol on how to perform peptide-centric SEC-SWATH-MS data analysis is available on GitHub at https://github.com/CCprofiler/SECSWATH_PeptideCentricAnalysis

The R-package *CCprofiler* is available on GitHub at https://github.com/CCprofiler/CCprofiler/

A detailed protocol on how to perform complex-centric SEC-SWATH-MS data analysis with the *CCprofiler* package as well as example data of our HEK293 experiment are available on GitHub at https://github.com/CCprofiler/SECSWATH_ComplexCentricAnalysis and in the supplementary *CCprofiler* manual.

References

1. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).
2. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505 (2017).
3. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).
4. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **196**, 801 LP – 810 (2012).
5. Liu, X., Yang, W., Gao, Q. & Regnier, F. Toward chromatographic analysis of

interacting protein networks. *J. Chromatogr. A* **1178**, 24–32 (2008).

6. Dong, M. *et al.* A “Tagless” Strategy for Identification of Stable Protein Complexes Genome-wide by Multidimensional Orthogonal Chromatographic Separation and iTRAQ Reagent Tracking. *J. Proteome Res.* **7**, 1836–1849 (2008).
7. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* **9**, 907 (2012).
8. Kristensen, A. R. & Foster, L. J. Protein Correlation Profiling-SILAC to Study Protein-Protein Interactions. in 263–270 (Humana Press, New York, NY, 2014). doi:10.1007/978-1-4939-1142-4_18
9. Havugimana, P. C. *et al.* A Census of Human Soluble Protein Complexes. *Cell* **150**, 1068–1081 (2012).
10. Wan, C. *et al.* Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015).
11. Kirkwood, K. J., Ahmad, Y., Larance, M. & Lamond, A. I. Characterization of Native Protein Complexes and Protein Isoform Variation Using Size-fractionation-based Quantitative Proteomics. *Mol. Cell. Proteomics* **12**, 3851–3873 (2013).
12. Larance, M. *et al.* Global Membrane Protein Interactome Analysis using *In vivo* Crosslinking and Mass Spectrometry-based Protein Correlation Profiling. *Mol. Cell. Proteomics* **15**, 2476–2490 (2016).
13. Scott, N. E. *et al.* Interactome disassembly during apoptosis occurs independent of caspase cleavage. *Mol. Syst. Biol.* **13**, 906 (2017).
14. Stacey, R. G., Skinnider, M. A., Scott, N. E. & Foster, L. J. A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). *BMC Bioinformatics* **18**, 457 (2017).
15. Heusel, M. *et al.* Complex-centric proteome profiling by SEC-SWATH-MS. *Mol. Syst. Biol.* **15**, e8438 (2019).
16. Scott, N. E., Brown, L. M., Kristensen, A. R. & Foster, L. J. Development of a computational framework for the analysis of protein correlation profiling and spatial proteomics experiments. *J. Proteomics* **118**, 112–129 (2015).
17. Pauling, L., Itano, H. A., Singer, S. J. & Wells, I. C. Sickle Cell Anemia, a Molecular Disease. *Science (80-.).* **110**, 543–548 (1949).
18. Bache, N. *et al.* A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).
19. Wessels, H. J. C. T. *et al.* LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. *Proteomics* **9**, 4221–4228 (2009).
20. Hu, L. Z. *et al.* EPIC: software toolkit for elution profile-based inference of protein complexes. *Nat. Methods* **16**, 737–742 (2019).
21. Glatter, T., Wepf, A., Aebersold, R. & Gstaiger, M. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.* **5**, 237 (2009).
22. Roncagalli, R. *et al.* Quantitative proteomics analysis of signalosome dynamics in primary T cells identifies the surface receptor CD6 as a Lat adaptor–independent TCR signaling hub. *Nat. Immunol.* **15**, 384–392 (2014).
23. Collins, B. C. *et al.* Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nat. Methods* **10**, 1246–1253 (2013).
24. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4**, 222 (2008).
25. Picotti, P. & Aebersold, R. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).
26. Schubert, O. T. *et al.* Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10**, 426–441 (2015).
27. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291 (2017).
28. Bruderer, R. *et al.* Optimization of Experimental Parameters in Data-Independent

Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol. &amp;amp; Cell. Proteomics* **16**, 2296 LP – 2309 (2017).

29. Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
30. Meier, F. *et al.* Parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF): Bottom-up proteomics with near optimal ion usage. *bioRxiv* 656207 (2019). doi:10.1101/656207
31. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 140031 (2014).
32. Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494**, 266–270 (2013).
33. Blattmann, P. *et al.* Generation of a zebrafish SWATH-MS spectral library to quantify 10,000 proteins. *Sci. Data* **6**, 190011 (2019).
34. Heusel, M. Complex-centric Proteome Profiling by SEC-SWATH Mass Spectrometry. (2017). doi:10.3929/ETHZ-B-000220300
35. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. &amp;amp; Cell. Proteomics* **11**, O111.016717 (2012).
36. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219 (2014).
37. Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **8**, 430–435 (2011).
38. Teleman, J. *et al.* DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* **31**, 555–562 (2015).
39. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* (2017). doi:10.1038/nmeth.4398
40. Röst, H. L. *et al.* TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13**, 777 (2016).
41. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2009).
42. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2012).
43. Szklarczyk, D. *et al.* STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
44. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–5 (2003).
45. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. in 339–368 (Humana Press, New York, NY, 2017). doi:10.1007/978-1-4939-6747-6_23
46. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
47. Hirano, Y. *et al.* A heterodimeric complex that promotes the assembly of mammalian 20S proteasomes. *Nature* **437**, 1381–1385 (2005).
48. Hirano, Y. *et al.* Dissecting β-ring assembly pathway of the mammalian 20S proteasome. *EMBO J.* **27**, 2204–2213 (2008).

Figure captions

Figure 1. Schematic overview of the complex-centric proteome profiling workflow. The workflow consists of three main modules: (1) Extraction of protein complexes from a biological sample and fractionation by size exclusion chromatography (SEC), (2) bottom-up proteomics analysis of all sampled fractions by data-independent acquisition mass

spectrometry and targeted, peptide-centric analysis (SWATH-MS), and (3) Inference of proteome assembly state and detection of specific protein complexes by targeted, complex-centric analysis within the R software package CCprofiler. The protocol takes as input a biological sample, requires some standard proteins for MW calibration, a spectral library for peptide-centric SWATH-MS data analysis and prior protein connectivity information. Output of the workflow are: (i) a quantitative assessment of the overall global proteome assembly state of the proteome analyzed, (ii) a quantitative assessment how each protein partitions into a certain number of SEC-resolvable distinct protein assembly states as well as (iii) a quantitative assessment on the protein complexes and sub-complexes in the given proteome at its current biological state.

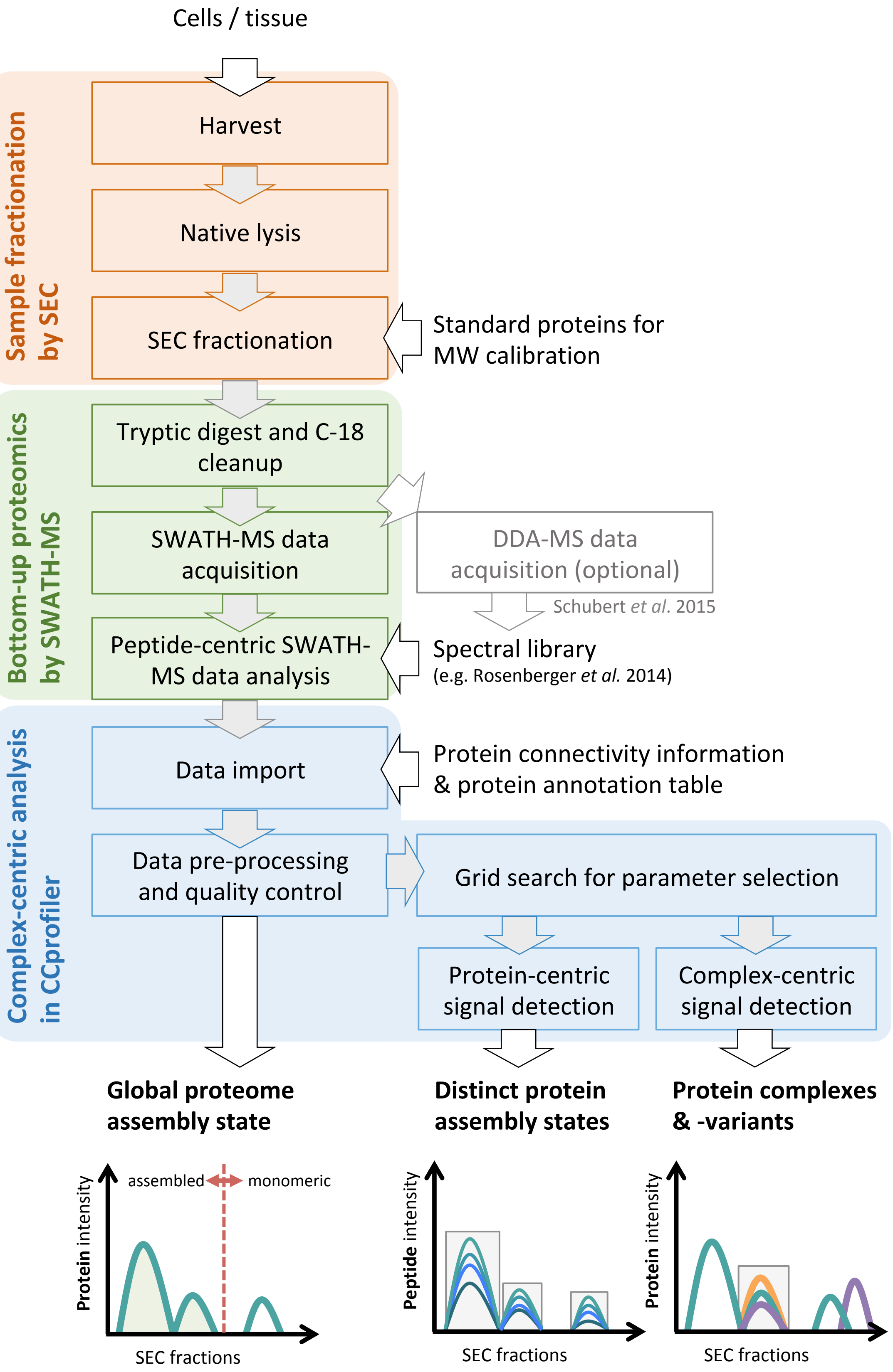
Figure 2. QC plots for SEC fractionation and SWATH-MS data acquisition. (A) The OD280 profile along the sampled SEC fractions for two replicates R1 and R2. Fractions collected from two consecutive, well reproducible separations (injections 1 and 2) of 1000 ug mild lysate each were pooled for downstream analyses. The first fraction that was measured by mass spectrometry for replicate R1 was labeled as fraction 1. Any prior fractions are labeled ≤0. **(B)** The total ion current (TIC) profiles of the DIA/SWATH-MS runs.

Figure 3. Data preprocessing plots. (A) Molecular weight calibration based on measured standard proteins and their molecular weights were: Thyroglobulin tetramer, 1398 kDa; Thyroglobulin dimer, 699 kDa; IgA, 300 kDa; IgG, 150 kDa; Ovalbumin, 44 kDa; and Myoglobin, 17 kDa. **(B)** Distribution of sibling-peptide correlations for both target proteins (solid line) and decoy proteins (dashed line). **(C)** Pseudo-ROC curves illustrating the effect of using the sibling-peptide correlation to perform FDR filtering. **(D)** Global statistics of protein signal attribution to assembled or monomeric state. The majority of detected protein mass (55%), as estimated by the total MS signal intensity, appears in assembled state in SEC-SWATH-MS. **(E)** Elution profiles of all peptides detected for the proteasome subunit alpha type-1 (PSA1) protein. The red vertically dashed line indicates the expected monomer MW. The salmon colored line indicates the selected cutoff for diving the elution range in assembled vs. monomeric, lying at twice the expected monomer MW.

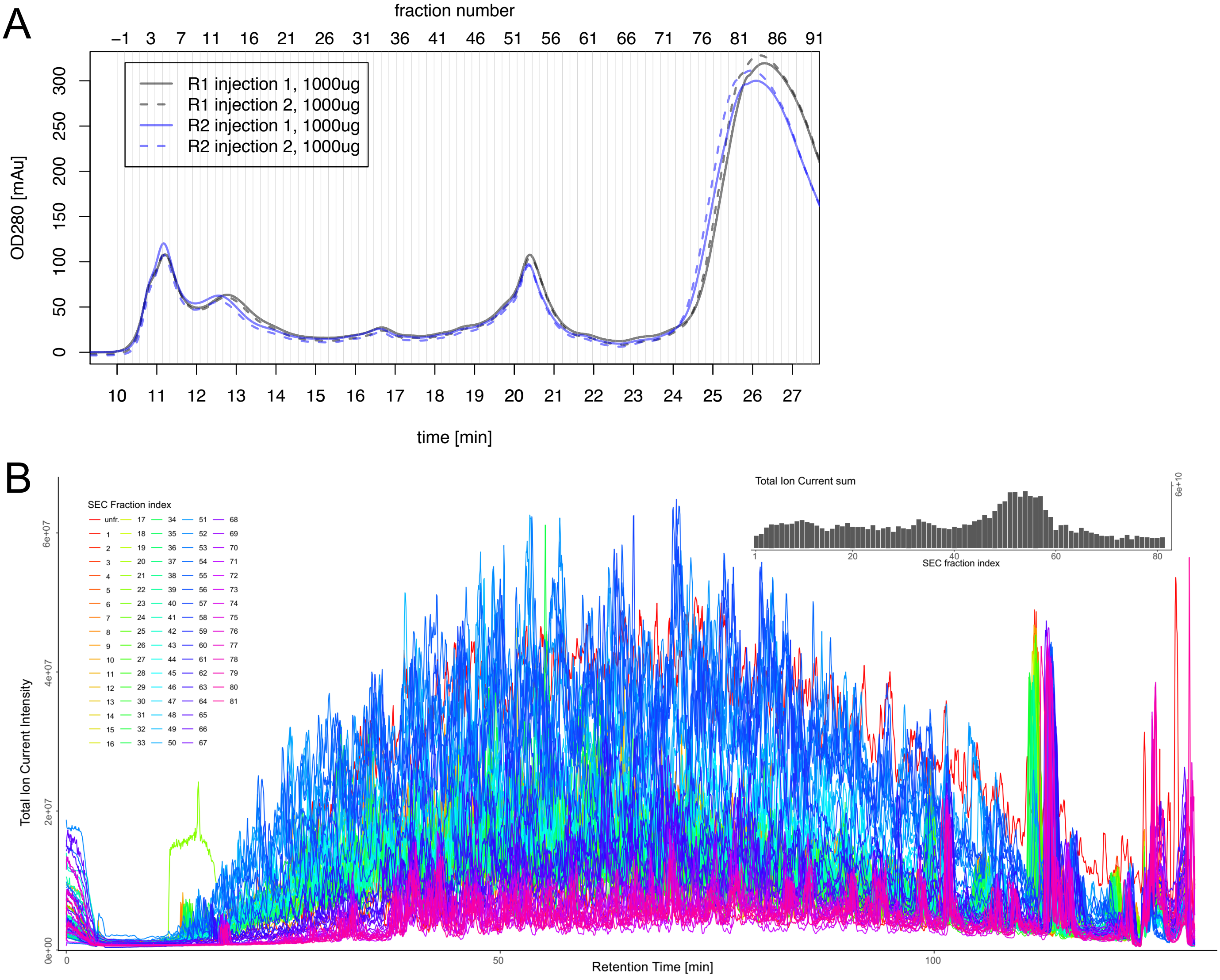
Figure 4. Parameter selection and protein-centric analysis. (A) Pseudo-ROC curves showing the number of estimated true positive protein features over increasing q-value (~ FDR) cutoffs for all tested parameter combinations. Here, each parameter set is colored according to the tested correlation cutoff. **(B)** P-value histogram for the protein-centric signal detection. **(C)** Pseudo-ROC curve showing the number of estimated true positive protein features over increasing q-value (~ FDR) cutoffs. **(D)** Histogram showing the number of proteins detected to elute in between one and seven distinct elution peaks. The pie chart illustrates that the majority of detected protein elution signals elute in the assembled MW range. **(E)** Elution profiles of all peptides detected for the proteasome subunit alpha type-1 (PSA1) protein. The protein elution signals determined by CCprofiler are highlighted in grey shading; peak apexes (solid) and boundaries (dashed) are shown as grey vertical lines. The red vertically dashed line indicates the expected monomer MW.

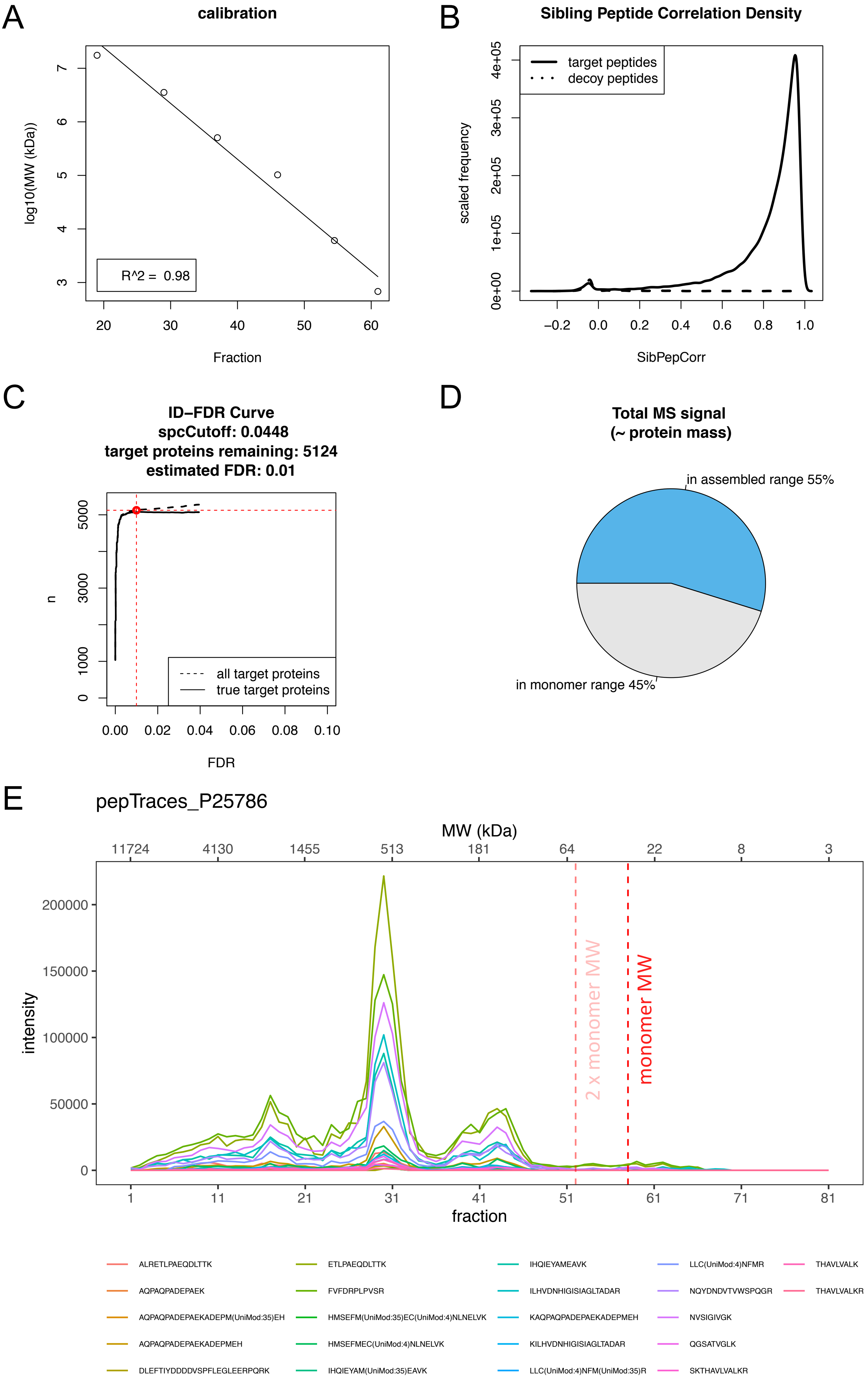
Figure 5. Exemplary CCprofiler plots for complex-centric data analysis. (A) Pie chat illustrating the number of proteins that are annotated in the CORUM protein complex database (2’532) and which fraction was confidently detected in the SEC-SWATH-MS data (1479). **(B)** P-value histogram for the complex-centric signal detection. **(C)** Pseudo-ROC curve showing the number of estimated true positive protein complex features over increasing q-value (~ FDR) cutoffs. **(D)** Histogram illustrating a summary of how many sub-complex signals were detected per protein complex query. The pie chart summarizes how many protein complexes annotated in the CORUM protein complex database have been detected by CCprofiler in the SEC-SWATH-MS data. **(E)** Protein elution profiles of the 22 protein subunits annotated to belong to the 26S proteasome complex. The protein-complex signals determined by CCprofiler are highlighted in grey shading; peak apex (solid) and boundaries

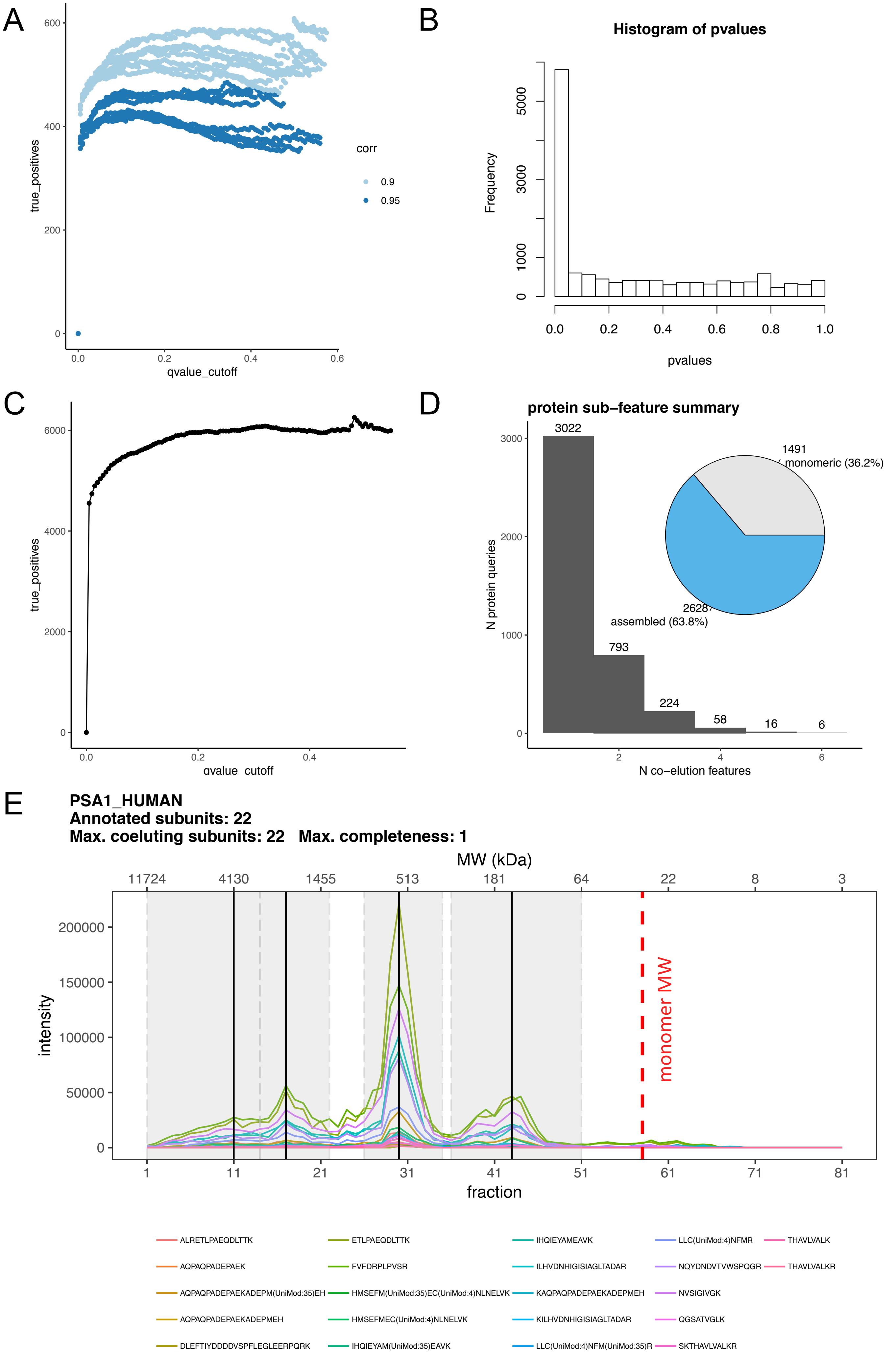
(dashed) are shown as grey vertical lines. Small triangles along the MW axis indicate the expected position of each of the subunits monomers.

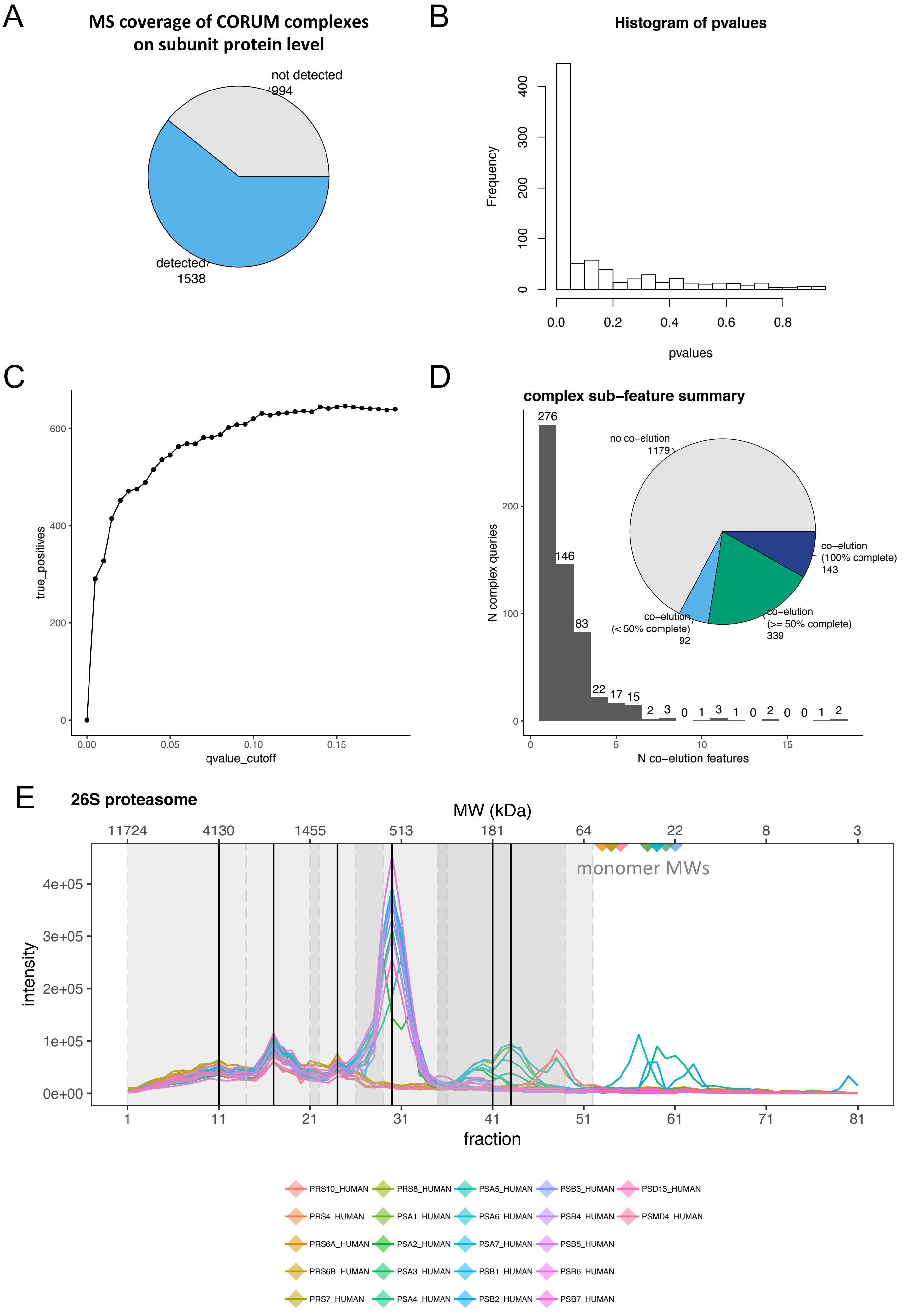


Cells / tissue
Sample fractionation by SEC
Harvest
Native lysis
SEC fractionation
Standard proteins for MW calibration
Bottom-up proteomics by SWATH-MS
Tryptic digest and C-18 cleanup
SWATH-MS data acquisition
DDA-MS data acquisition (optional)
Schubert et al. 2015
Peptide-centric SWATH-MS data analysis
Spectral library
(e.g. Rosenberger et al. 2014)
Complex-centric analysis in CCprofiler
Data import
Protein connectivity information & protein annotation table
Data pre-processing and quality control
Grid search for parameter selection
Protein-centric signal detection
Complex-centric signal detection
Global proteome assembly state
Distinct protein assembly states
Protein complexes & -variants
Protein intensity
assembled
monomeric
SEC fractions
Peptide intensity
SEC fractions
Protein intensity
SEC fractions


A
fraction number
−1 3 7 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91
OD280 [mAu]
0 50 100 150 200 250 300
R1 injection 1, 1000ug
R1 injection 2, 1000ug
R2 injection 1, 1000ug
R2 injection 2, 1000ug
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
time [min]
B
SEC Fraction index
unfr. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67
68 69 70 71 72 73 74 75 76 77 78 79 80 81
Total Ion Current Intensity
0e+00 2e+07 4e+07 6e+07
0 50 100
Retention Time [min]
Total Ion Current sum
6e+10
1 20 40 60 80
SEC fraction index


A
calibration
log10(MW (kDa))
7
6
5
4
3
R^2 = 0.98
20
30
40
50
60
Fraction
B
Sibling Peptide Correlation Density
scaled frequency
0e+00
1e+05
2e+05
3e+05
4e+05
target peptides
decoy peptides
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
SibPepCorr
C
ID−FDR Curve
spcCutoff: 0.0448
target proteins remaining: 5124
estimated FDR: 0.01
n
0
1000
3000
5000
all target proteins
true target proteins
0.00
0.02
0.04
0.06
0.08
0.10
FDR
D
Total MS signal
(~ protein mass)
in assembled range 55%
in monomer range 45%
E
pepTraces_P25786
MW (kDa)
11724
4130
1455
513
181
64
22
8
3
intensity
200000
150000
100000
50000
0
2 x monomer MW
monomer MW
1
11
21
31
41
51
61
71
81
fraction
ALRETLPAEQDLTTK
AQPAQPADEPAEK
AQPAQPADEPAEKADEPM(UniMod:35)EH
AQPAQPADEPAEKADEPMEH
DLEFTIYDDDDVSPFLEGLEERPQRK
ETLPAEQDLTTK
FVFDRPLPVSR
HMSEFM(UniMod:35)EC(UniMod:4)NLNELVK
HMSEFMEC(UniMod:4)NLNELVK
IHQIEYAM(UniMod:35)EAVK
IHQIEYAMEAVK
ILHVDNHIGISIAGLTADAR
KAQPAQPADEPAEKADEPMEH
KILHVDNHIGISIAGLTADAR
LLC(UniMod:4)NFM(UniMod:35)R
LLC(UniMod:4)NFMR
NQYDNDVTVWSPQGR
NVSIGIVGK
QGSATVGLK
SKTHAVLVALKR
THAVLVALK
THAVLVALKR


A
600
400
200
0
true_positives
0.0
0.2
0.4
0.6
qvalue_cutoff
corr
0.9
0.95
B
Histogram of pvalues
Frequency
0
1000
3000
5000
0.0
0.2
0.4
0.6
0.8
1.0
pvalues
C
6000
4000
2000
0
true_positives
0.0
0.2
0.4
qvalue_cutoff
D
protein sub−feature summary
N protein queries
3000
2000
1000
0
3022
793
224
58
16
6
2
4
6
N co−elution features
1491
monomeric (36.2%)
2628
assembled (63.8%)
E
PSA1_HUMAN
Annotated subunits: 22
Max. coeluting subunits: 22 Max. completeness: 1
MW (kDa)
11724
4130
1455
513
181
64
22
8
3
intensity
200000
150000
100000
50000
0
monomer MW
1
11
21
31
41
51
61
71
81
fraction
ALRETLPAEQDLTTK
AQPAQPADEPAEK
AQPAQPADEPAEKADEPM(UniMod:35)EH
AQPAQPADEPAEKADEPMEH
DLEFTIYDDDDVSPFLEGLEERPQRK
ETLPAEQDLTTK
FVFDRPLPVSR
HMSEFM(UniMod:35)EC(UniMod:4)NLNELVK
HMSEFMEC(UniMod:4)NLNELVK
IHQIEYAM(UniMod:35)EAVK
IHQIEYAMEAVK
ILHVDNHIGISIAGLTADAR
KAQPAQPADEPAEKADEPMEH
KILHVDNHIGISIAGLTADAR
LLC(UniMod:4)NFM(UniMod:35)R
LLC(UniMod:4)NFMR
NQYDNDVTVWSPQGR
NVSIGIVGK
QGSATVGLK
SKTHAVLVALKR
THAVLVALK
THAVLVALKR


A
MS coverage of CORUM complexes on subunit protein level
not detected 994
detected 1538
B
Histogram of pvalues
Frequency
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
pvalues
C
true_positives
0
200
400
600
0.00
0.05
0.10
0.15
qvalue_cutoff
D
complex sub−feature summary
N complex queries
0
100
200
276
146
83
22
17
15
2
3
0
1
3
1
0
2
0
0
1
2
0
5
10
15
N co−elution features
no co−elution 1179
co−elution (100% complete) 143
co−elution (>= 50% complete) 339
co−elution (< 50% complete) 92
E
26S proteasome
MW (kDa)
11724
4130
1455
513
181
64
22
8
3
monomer MWs
intensity
0e+00
1e+05
2e+05
3e+05
4e+05
1
11
21
31
41
51
61
71
81
fraction
PRS10_HUMAN
PRS4_HUMAN
PRS6A_HUMAN
PRS6B_HUMAN
PRS7_HUMAN
PRS8_HUMAN
PSA1_HUMAN
PSA2_HUMAN
PSA3_HUMAN
PSA4_HUMAN
PSA5_HUMAN
PSA6_HUMAN
PSA7_HUMAN
PSB1_HUMAN
PSB2_HUMAN
PSB3_HUMAN
PSB4_HUMAN
PSB5_HUMAN
PSB6_HUMAN
PSB7_HUMAN
PSD13_HUMAN
PSMD4_HUMAN