bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SECAT: Quantifying differential protein-protein interaction states by network-centric analysis

George Rosenberger[1], Moritz Heusel[2,3], Isabell Bludau[2,4], Ben Collins[2,5,#], Claudia Martelli[2,#], Evan Williams[2,#], Peng Xue[2,6,#] Yansheng Liu[7,8,#] Ruedi Aebersold[2,9] Andrea Califano[1,10,11,12,13]

1 Department of Systems Biology, Columbia University, New York, NY, USA
2 Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
3 Present address: Department of Clinical Sciences, Lund, Division of Infection Medicine, Lund University, Lund, Sweden
4 Present address: Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
5 Present address: School of Biological Sciences, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom
6 Institute of Biophysics, Chinese Academy of Sciences, Beijing, China
7 Yale Cancer Biology Institute, Yale University, West Haven, CT, USA
8 Department of Pharmacology, Yale University School of Medicine, New Haven, CT, USA
9 Faculty of Science, University of Zurich, Zurich, Switzerland
10 Department of Biomedical Informatics, Columbia University, New York, NY, USA
11 Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA
12 J.P. Sulzberger Columbia Genome Center, Columbia University, New York, NY, USA
13 Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA
 # Authors listed alphabetically

Correspondence to: aebersold@imsb.biol.ethz.ch and ac2248@cumc.columbia.edu

Abstract

Protein-protein interactions (PPIs) play a critical role in virtually all cellular processes. Their context-dependent characterization is thus a key objective of proteomic research. We and others have previously shown that chromatographic fractionation of native protein complexes (e.g. through size-exclusion chromatography, SEC) can be effectively combined with high-throughput, bottom-up mass-spectrometry-based proteomics (e.g. data-independent acquisition-based SWATH-MS), to support proteome-wide characterization of protein complexes.

To enable qualitative and quantitative comparison of the proteome organization encoded in these datasets, across multiple experimental conditions, scalable and robust

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analysis strategies are required. To address this need, we developed the Size-Exclusion Chromatography Algorithmic Toolkit (SECAT), a novel network-centric strategy for the quantitation of protein complex profiles. SECAT elucidates proteins and their context-specific PPIs in terms of both abundance and connectivity. We validate algorithm predictions using publicly available datasets and compare them to established strategies to demonstrate that SECAT represents a more scalable and effective methodology to assess protein-network state, obviating the need to infer individual protein complexes. Further, by comparing PPI-networks in interphase and mitotic HeLa cells, we demonstrate SECAT's ability to provide novel insight about context-specific molecular mechanisms that differentiate cellular states.

Introduction

Living cells depend on a large number of coordinated and concurrent biochemical reactions. Most of these are catalyzed and controlled by macromolecular entities of well-defined subunit composition and 3D structure, a notion that has been captured by the term "modular cell biology" by Hartwell and colleagues [1]. The vast majority of these modules consist of or contain protein complexes. It is thus a basic assumption of the modular cell biology model that alterations in protein complex structure, composition and abundance will alter the biochemical state of cells. Elucidating protein complexes and their organization in extended protein-protein interaction (PPI) networks is, therefore, of paramount importance for both basic and translational research.

Traditionally, the composition and structure of protein complexes has been determined by two broad, complementary approaches, structural biology and interaction proteomics. Structural biology encompasses a suite of powerful tools to characterize individual, purified or reconstituted protein complexes at high resolution, including at the individual atomic coordinates level. High resolution structures have provided a wealth of functional and mechanistic insights into biochemical reactions [2]. However, they have been solved for only a few hundred human protein complexes and the number of cases where the structure of protein complexes is assessed across different functional states is even lower. This is contrasted by the observation that in protein cell extracts, under mild lysis conditions, approximately 60% of proteins and total protein cell mass is engaged in protein complexes [3]. As a result, methodologies for the rapid elucidation of protein complexes are still critically needed.

Interaction proteomics encompasses a suite of methods to determine composition and, at times, subcellular location and abundance of protein complexes, albeit at lower resolution but higher throughput than structural biology techniques. Among these methods, liquid chromatography coupled to tandem mass spectrometry [4, 5] (LC-MS/MS)—and more specifically affinity purification coupled to LC-MS/MS (AP-MS [6])—has been most widely used. For AP-MS analysis, individual proteins are engineered to display an affinity tag and are expressed as "bait" proteins in cells. The bait and the corresponding "prey" proteins assembled around it are then isolated and qualitatively [7–10] or quantitatively [11–14] analyzed by MS. This method has proven robust across laboratories [15] and, through process automation and integrative data analysis, efforts to map the PPIs of the entire human proteome [16, 17] have helped characterize more than half of canonical human proteins [18]. This knowledge is embedded in a variety of databases, including BioPlex [16, 18], STRING [19], IntAct [20] and hu.MAP [21] that provide access to generic human PPI maps. They have also been used to predict PPIs for previously uncharacterized proteins, e.g. PrePPI [22–24]. Whereas it can be expected that these systematic PPI mapping projects will reach

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

saturation in the next few years, AP-MS and related approaches are fundamentally limited in their ability to detect compositional or abundance changes across multiple cell states and to detect concurrent changes in different complexes within the same sample.

Protein correlation profiling (PCP) [25] and related methods [3, 26–31] have been proposed as a means to concurrently analyze multiple complexes from the same sample. PCP proceeds by first separating native protein complexes—e.g. according to their hydrodynamic radius by size-exclusion chromatography (SEC)—then collecting 40-80 consecutive fractions, and finally performing bottom-up mass spectrometric analysis of fraction-specific proteins. The result is a set of quantitative protein abundance profiles, across the apparent molecular weight of each protein complex (Fig. 1a). Under the assumption that protein subunits of the same complex have congruent SEC profiles, the latter can then be used to infer protein-protein interactions and protein complex composition. Conditional to availability of quantitative mass spectrometric data, the method further supports comparative analysis across multiple biological conditions, which is critically missing in current PPI databases. For the most part, PCP datasets have been analyzed using interaction-centric algorithms [26, 27, 29, 31–34] which essentially use chromatographic co-elution of protein profiles to identify PPIs, to infer protein complexes and to conduct qualitative and quantitative comparisons across biological conditions. Interaction-centric algorithms are limited by the inherently low SEC resolution, resulting in the presence of hundreds to thousands of proteins per SEC fraction. This lowers the confidence of inferred interactions because of the high probability that non-interacting proteins may show similar elution profiles by chance [34].

To address these limitations, we recently developed the algorithm CCprofiler [3] which implements a targeted, complex-centric strategy to query the protein elution profiles generated by high resolution SEC-SWATH-MS [3] to assess presence, composition and abundance of predefined protein complexes. Similar to targeted proteomic approaches [35, 36], transforming the problem from de novo "identification" of unknown protein complexes to a posteriori "detection" of established ones significantly increases sensitivity. Using prior knowledge from reference databases like CORUM [37], BioPlex [16, 18] or STRING [19], the confidence of protein complex detection was also substantially improved, albeit at the cost of missing potential novel interactions or complexes not included in the query set. Taking advantage of the precise quantitative values generated by SWATH-MS, we have also shown that the CCprofiler strategy is well-suited to detect changes in complex-associated protein abundance across conditions [38].

However, several critical challenges remain, both for interaction- and complex-centric strategies. This is especially relevant in studies where an increasing number of experimental conditions or samples are compared, as can be expected in the near future. Specifically, the assumption that protein complexes constitute entities of fixed subunit composition across different contexts and conditions is both highly restrictive and biologically unrealistic, thus critically hindering data analysis and interpretation. For example, the spliceosome—a complex molecular machinery controlling pre-mRNA intron removal—consists of small nuclear RNAs (snRNA) and more than 100 protein subunits, assembled into a variety of submodules, each with highly context-specific activity and composition [39, 40]. Since only a fraction of these protein subunits will be detectable across all conditions in typical bottom-up proteomics experiments, differentiating between biological effects and technical artefacts is challenging for the absent subunits if complexes are considered to be static entities.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

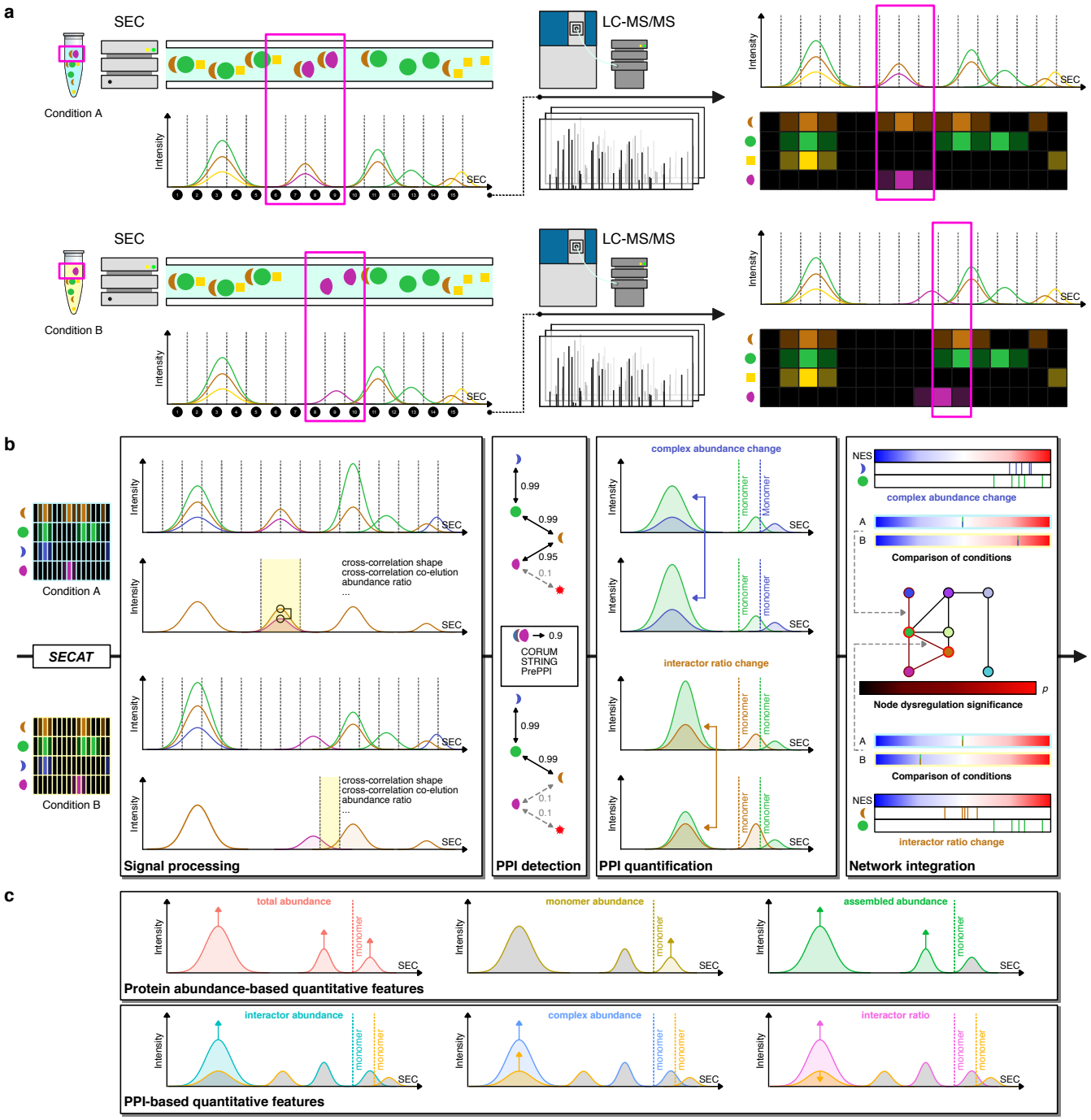


Figure 1. Methodological overview. (**a**) The SEC-SWATH-MS workflow is depicted comparing two different experimental conditions A and B. Native protein extracts are separated by SEC and individual fractions are sampled (1-15). Each fraction is separately measured by LC-MS/MS. Peptide or protein abundance grouped according to SEC fraction can be depicted as profile, whereby co-eluting proteins might be part of the same protein complex. Two experimental conditions can result in similar protein complex profiles but might differ for some proteins (highlighted in pink). (**b**) The SECAT algorithm uses peptide-level protein complex profiles and reference PPI networks as input. The signal processing module computes a set of scores for the peptide profiles of candidate interactors. Using the set of scores and ground truth positive and negative interactomes, the PPI detection module conducts semi-supervised

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

learning and generates context-specific PPI subnetworks. Confident PPIs are quantified to assess complex abundance and interactor ratio changes using the PPI quantification module. The proteoVIPER algorithm based on VIPER [41] is used for quantitative protein inference from peptide-level abundances. Network integration: Significance of changes between the experimental conditions is integrated by a network-centric approach on all different levels. (**c**) SECAT provides quantitative insights on different levels. Protein abundance-based quantitative metrics are computed for total abundance, monomer abundance and assembled abundance. PPI-based quantitative metrics are provided separately for each interactor, integrated as positively correlated complex abundance or negatively correlated interactor ratio metric.

To address these limitations, we introduce the Size-Exclusion Chromatography Algorithmic Toolkit (SECAT). This methodology further extends PCP data analysis to support network-centric analysis, without requiring the inference of protein complexes. In a first step, SECAT scores binary PPIs from PCP datasets to construct context-specific, error-rate-controlled PPI networks. In a second step, rather than using the resulting connectivity maps to directly infer protein complexes, SECAT leverages them to transform differential protein abundances across fractions into metrics representing complex abundance and interactor-protein ratio. We demonstrate that this novel, network-centric strategy for protein complex attribute inference is robust against technical variability and overcomes key limitations of current protein-complex inference methods, while representing changes in network connectivity in an intuitive and unbiased way. This supports discovery of cell-state-specific PPI network changes as well as the molecular mechanisms associated with them.

Results

The Size-Exclusion Chromatography Algorithmic Toolkit

SECAT's workflow consists of four consecutive steps: signal processing, PPI detection, PPI quantification and network integration, respectively. The main input data are the peptide-level PCP profiles, e.g. acquired by SEC-SWATH-MS [3,38], and a set of reference PPI networks (Fig. 1b, Methods). The output of the system are context-specific interactomes with protein-level metrics summarizing the associated differential protein and PPI properties (Fig. 1c, Methods). In the following we describe each step.

Step 1, Signal processing: The goal of the first step is to define the SEC profile boundaries within which specific binary interactions are observable and to derive a set of partial scores that discriminate true vs. false interactions. Candidate interactions are obtained from a comprehensive repertoire of PPI network references, such as CORUM [37], STRING [19] or PrePPI [24]. Since protein complexes frequently form Gaussian elution peaks in the SEC dimension, previous approaches [3,31,34] preprocessed the data to select peak boundaries and to filter the data for the most promising candidate signals ("peak picking"). In contrast, to maintain quantitative consistency between conditions and replicates, SECAT avoids peak-picking altogether and performs only minimal signal preprocessing, thus focusing the analysis on the SEC fraction subsets where both tested interactors are detectable (Methods). For each candidate PPI, peptide-level elution profiles from each protein are processed to compute chromatographic (cross-correlation shape and shift [42,43], maximal and total information criterion [44]), interactor ratio, SEC coverage, and monomeric fraction distance metrics (Fig. 1b, Methods). The result of this step is a table in which each candidate PPI is associated with a score vector representing the different properties that can differentiate between true and false interactions.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Step 2, PPI detection: To generate context-specific interaction subnetworks, partial scores from the first step are used as input to a machine learning (ML) approach based on the PyProphet algorithm [45, 46]. This step is designed to discriminate true vs. false interactions and to estimate their confidence. For this purpose, SECAT is trained using a set of true negative PPIs as null model and the most confidently detected PPIs as gold standard model of true interactions (Fig. 1b, Methods). Since queried or tested candidate PPIs may, by chance, match co-eluting protein profiles that are not representative of true positive interactions detectable within the experimental context, SECAT uses a semi-supervised learning approach (Fig. 1b, Methods). The first iteration is thus initialized by a composite score only for the most confident PPIs, with subsequent iterations increasing the classifier sensitivity, relying on cross-validation and "early stopping" to alleviate overfitting (Methods). Many reference PPI networks provide confidence metrics for individual interactions; SECAT can incorporate the prior confidence by computing a group-based false discovery rate (FDR) metric (Methods). Thus, SECAT learns a classifier using only high confidence PPI networks (e.g. CORUM [37]) and then applies it to less stringent networks (e.g. STRING [19] or PrePPI [24]) to maximize sensitivity and coverage of profiled interactions. In summary, this step generates a set of confidence metrics for each candidate PPI (posterior error probabilities and q-values), that allow thresholding the list at any user-defined FDR.

Step 3, PPI quantification: For each protein involved in an interaction, the molecular mass of its monomeric subunit is used, together with the SEC calibrated molecular weight scale [3] to define a protein-specific "monomer threshold", which corresponds to the SEC fraction index that separates the fractionation range into regions where the protein of interest is present in likely assembled or monomeric forms (Methods). Within the fractions covering the assembled protein conformations, the combined set of PPIs that are confidently detected across all experimental conditions and replicates is then used for quantification. Specifically, for each protein in a candidate binary interaction, peptide-level data within the boundaries of the SEC fractions defined in Step 1 are independently summarized. This provides quantitative metrics that can be used to assess protein or PPI changes across experimental conditions (Fig. 1b, Methods).

Step 4, Network integration: The goal of the last step is to integrate individual protein and binary PPI metrics from the previous step into comprehensive PPI network states, where nodes represent individual proteins and their differential attributes and edges indicate specific binary PPIs, representing the consensus across different conditions and replicates (Fig. 1b, Methods).

The VIPER [41] algorithm was originally developed to infer protein activity from target transcript abundances using gene regulatory networks (GRN). SECAT's proteoVIPER algorithm extends this conceptual strategy by adapting it to proteomic data that estimates protein- or PPI-level metrics, based on their fractionated peptide abundances, as computed by the previous step (Methods). To estimate protein-level metrics, peptide-protein relationships are used to compute a normalized enrichment score (NES). When all fractions are considered, a metric for "total" protein abundance can thus be computed; alternatively, considering the above-defined monomer threshold, metrics for "monomeric" (right side of threshold) or "assembled" (left side of threshold) subunits can be estimated (Fig. 1c, Methods). To estimate PPI-level metrics, the peptides of two interactor proteins are quantified using only their overlapping SEC fractions. For each protein, a separate "interactor" metric can be computed as described above to assess the quantitative changes for the individual protein within the particular PPI. Alternatively, complex-related metrics can be derived: First, the peptides of the two interactors are assessed in a positively correlated setting, where the

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

resulting metric can be used to assess "complex abundance". Second, the interactor peptides are assessed in a negatively correlated setting to derive a metric representing "interactor ratio" (Fig. 1c, Methods). This metric can represent stoichiometric changes between interacting subunits within a complex or alterations in their connectivity, if a PPI is abrogated in some conditions.

In the final step, SECAT statistically compares the quantitative metrics between different groups and then integrates PPI-level (edges) metrics on a protein-by-protein basis (nodes) to assess protein abundance or protein-complex-based changes for every protein of interest. For this integration, we use the Empirical Brown's Method [47], which is specifically designed to account for non-statistically-independent evidence and can integrate both redundant (PPI for protein of interest from the same complex) and orthogonal (PPI for protein of interest from different complexes) information (Methods). In conclusion, SECAT leverages large-scale SEC-SWATH-MS measurements, including across multiple conditions, to generate context-specific PPI networks, as well as protein abundance and other PPI-related quantitative metrics, to comprehensively characterize condition-specific proteome and proteome-network changes.

Parameter selection and validation of signal processing and PPI detection modules

To assess SECAT's signal processing and PPI detection modules for parameter selection and validation, we first benchmarked the effect of different peak-picking preprocessing methods. For this, we used a publicly available SEC-SWATH-MS dataset [38] of a HeLa CCL2 cell line that was measured in triplicate in both interphase and mitotic cell state (referred to as HeLa-CC dataset, Methods). Then, we assessed the method's robustness based on its ability to infer bona fide PPIs using reference networks with an increasing ratio of false vs. true PPIs. For both assessments, the entire SECAT methodology was used, however, with different choices for the parameters at each step.

Defining reference sets of true and false PPIs is non-trivial and for this benchmarking process we adapted a previously applied strategy used by the PrInCE algorithm [34] (Methods). It leverages CORUM [37] PPIs as true and all other interactions of CORUM proteins that are not included in the database (CORUM-inverted) as false PPIs. Since proteins in CORUM complexes are well characterized and, for the most part, supported by 3D structure data, this strategy assumes that any true interactions within that set of proteins should be known already. As such, identified or detected interactions that are not reported are likely false positives [34]. For the purpose of the benchmark we further excluded any known or predicted interactions from CORUM-inverted and split the combined true/false reference set into equally sized training/validation and hold-out subsets (Methods).

Effect of signal preprocessing on PPI detection consistency: Both, interaction- and complex-centric algorithms frequently employ peak picking to increase signal-to-noise (S/N) characteristics of co-eluting candidate peaks. This can lower detection consistency due to the often-extreme elution profile heterogeneity of protein complex elution peaks (i.e., their variability in elution width from 5-60 SEC fractions), convoluted peak structure, intensities spanning several orders of magnitude, and different S/N characteristics. As a result, peaks may remain undetected—or different (elution) boundaries may be set across conditions—thus impairing resulting accuracy and reproducibility. To minimize these effects, SECAT skips peak picking and focuses only on the intersecting fractions of candidate interactors, as described in Step 1. To better assess the effects of this strategy, we compared it to two different peak picking

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

strategies, termed "detrend" (base line removal) and "localmax" (local maximum peak picking) (Methods).

Using the above-described benchmark setup, we first processed three replicates of each condition of the HeLa-CC dataset with the different peak picking options. Our results (Fig. 2a) show that peak-picking based noise removal results in tighter and less variable mean peak widths across all conditions and replicates. However, when comparing peak-width standard deviation across the three replicates, the "localmax" method introduced slightly higher variability (Fig. 2a). In contrast, the quantitative, integrated peak area-based metrics and particularly the standard deviation are very similar across all three options (Fig. 2a). This suggests that the three approaches perform similarly in terms of quantitative applications. However, considerable differences were found when comparing detection consistency (Fig. 2b). Indeed, decomposing the total number of detected PPIs for detectability across experimental replicates revealed that the fractions of PPIs consistently detected across replicates was substantially higher if no peak picking preprocessing was used, thus supporting its omission in SECAT.

Error-rate estimation accuracy and PPI detection sensitivity: To optimally separate true vs. false candidate PPIs, SECAT integrates 10 different partial scores into a composite score (Fig. 2c), using a semi-supervised learning strategy initialized by a high confidence PPI score (kickstart_score, Methods). Consistent with other evidence integration approaches, the final integrative score (secat_score) is more discriminative than any of the initial or partial scores (Fig. 2c). To assess error-rate estimation accuracy and algorithm sensitivity, we performed cross-validation analysis by applying the classifier to the hold-out subset with increasing fractions (1:0 –1:16) of false reference interactions (Methods, Fig. 2d). The accuracy of PPI detection confidence was then assessed by comparing the q-value estimates with the ground truth estimated FDR (Fig. 2e). Our results show that the q-value estimates for the hold-out dataset were accurate within the assessed range. This shows that the SECAT PPI detection module is robust against a variable fraction of false or undetectable PPIs in reference databases, especially in the more relevant high-confidence region (q-value < 0.1). However, lowering the specificity of the reference databases or hypothetically assessing all pairwise PPI combinations has a negative effect on the sensitivity due to multiple hypotheses testing correction [48] (Fig. 2f): Indeed, at a high confidence level (q-value < 0.05), only 1,173 bona fide PPIs were detected, using a 1:16 ratio of true/false PPIs in the query set, compared to 3,200, when a 1:0 ratio was used. The higher PPI recovery rate illustrates the benefit of using a high-quality reference PPI network, rather than comparing all potential PPI interactions using the proposed scoring approach.

Compatibility with different data modalities: PCP datasets have been acquired by proteomic methods different from SWATH-MS, e.g. by label-based [30] or label-free [27] data-dependent acquisition methods. We tested the performance of SECAT with a publicly available SILAC-PCP dataset [30] which compared Anti-Fas IgM treated (inducing apoptosis) against control samples of Jurkat cells in triplicates per condition. We refer to this dataset as Jurkat-Fas. The assessment was conducted using the same strategy as described above.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

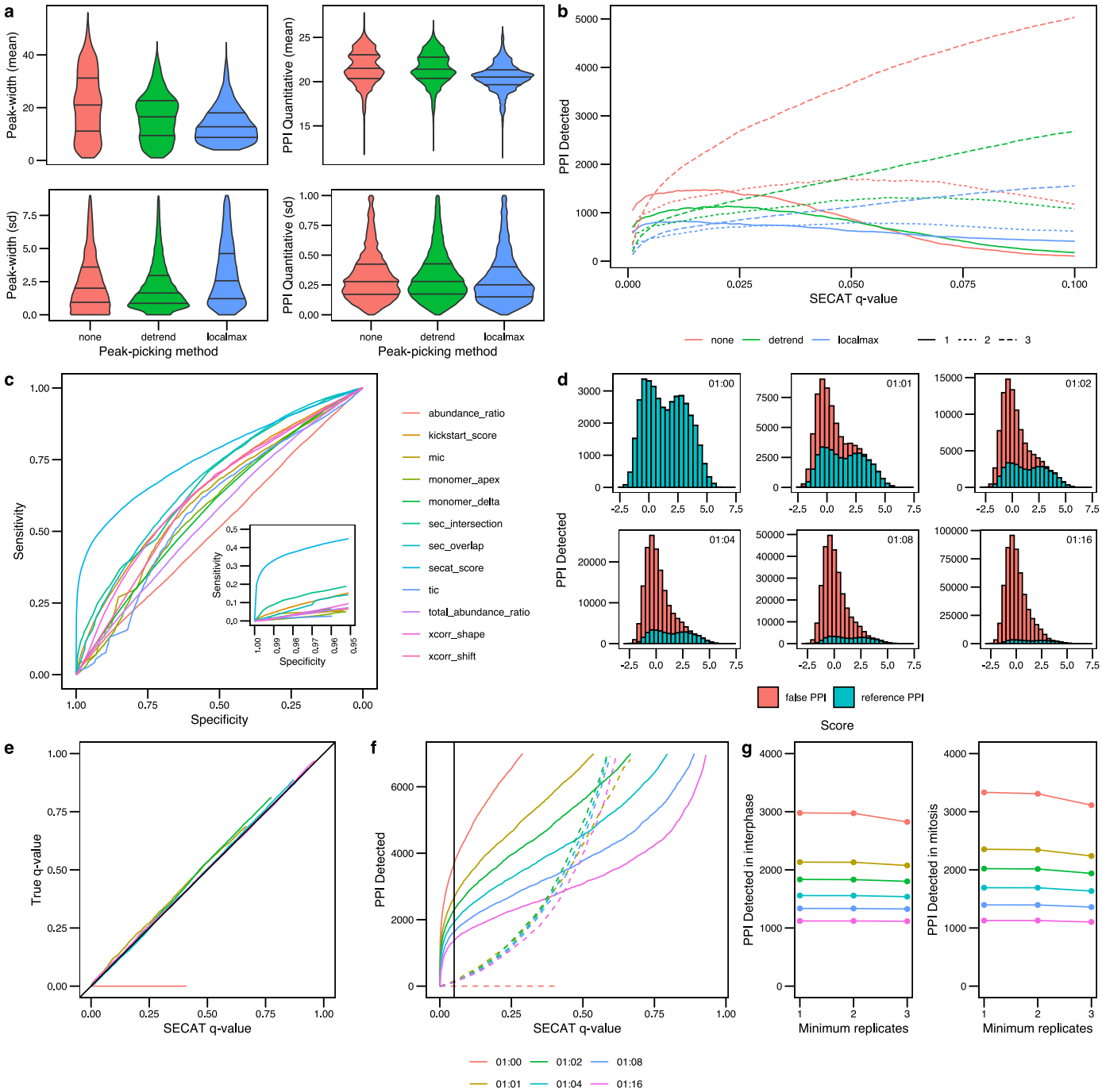


Figure 2. Signal processing and PPI detection performance evaluated using the HeLa-CC [38] dataset. (**a**) The effect of different peak-picking methods for signal processing on peak-width and PPI quantification within replicates of the same experimental condition is depicted in violin plots (lines representing 25, 50 and 75% quantiles respectively, Methods). (**b**) Sensitivity of PPI detection vs. SECAT q-value for peak-picking options. The PPI were filtered to a global context q-value < 0.05. Colors indicate different peak-picking options; line types indicate number of detections with the three replicates per condition. (**c**) The receiver-operating characteristic (ROC) illustrates the sensitivity vs. specificity of the different SECAT partial scores, the kickstart score and the integrated SECAT score. (**d**) The SECAT score histograms depict the dilution of the reference PPI network (true: CORUM) with false PPI (false:

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CORUM-inverted) for the benchmark (Methods). (**e**) The SECAT estimated q-value is accurate as evaluated against the ground truth for all levels of reference PPI network accuracy. For the dilution 1:0, no a priori false PPI are tested, thus the true FDR equals to 0. (**f**) The sensitivity of PPI detection in dependency of SECAT q-value for different levels of reference PPI network accuracy is depicted. Solid lines represent the true, whereas dashed lines represent the false PPI. (**g**) Consistency of PPI detection between replicates of the same conditions.

Benchmark data shows that SECAT can be effectively applied to the Jurkat-Fas dataset, with similar performance characteristics in terms of signal processing and error-rate control robustness, resulting in 2,049 PPIs (q-value < 0.05) detected across both conditions and replicates (Supplementary Fig. 1). A crucial requirement for studies comparing multiple experimental conditions and replicates is high PPI detection reproducibility. For example, requiring PPIs to be detected in all three instead of just one replicate reduced recall by 33.7 - 35.3% (Supplementary Fig. 1g) in the Jurkat-FAS dataset. In comparison, the corresponding drop for SEC-SWATH-MS in the HeLa-CC dataset was substantially smaller (5.2 - 6.6%) (Fig. 2g). In conclusion, the core assumptions for SECAT are also fulfilled for datasets acquired by different methods. But because different biological systems were investigated (i.e., HeLa-CC vs. Jurkat-Fas dataset), no conclusions can be made in terms of performance metric differences, since they could be of technical or biological (e.g. sample complexity) origin.

Comparison of SECAT with established algorithms: To assess the performance of the SECAT PPI detection module with that of established algorithms, we used a previously published SEC-SWATH-MS dataset of a HEK293 cell line [3] in exponential growth state, measured in a single replicate. In the following we refer to this dataset as HEK293-EG. We used the CORUM reference PPI network to generate a pseudo ground-truth dataset for classifier training, as described above, but restricted validation results to a manually curated list of detectable complexes (Methods).

To compare binary PPI identification/detection performance, we trained the SECAT, CCprofiler [3] and EPIC [31] classifiers using the same input data. For CCprofiler, complex hypotheses were defined as binary complexes. Since the algorithms apply very different signal processing and filtering steps, the numbers of reported PPIs with predicted scores varied substantially. SECAT only queries targeted PPIs and reported 39,288 PPIs of which 2,588 had a q-value < 0.05. Similarly, CCprofiler conducts targeted analysis, but the preprocessing steps of the algorithm reduced the number of reported PPIs to 5,092 of which 2,169 had a q-value < 0.05. EPIC assesses all combinations of protein interactions in an unbiased manner and only reported results with a score above 0.5 (not error-rate controlled), resulting in 4,678 PPIs.

To assess the power of individual classifiers to discriminate true and false PPIs, we compared all combinations of the three classifiers and also the full intersection of all predicted PPIs (without cutoffs, except EPIC score > 0.5). The full intersection of the three algorithms resulted in a set of predictions for 2,647 ground truth "true" and 222 ground truth "false" PPIs. On this set, SECAT achieved an area under the receiver operating characteristic (AUROC) of 0.835, with CCprofiler reporting a similar performance of 0.819, whereas the AUROC of EPIC was lower at 0.707 (Supplementary Fig. 2a). Further, SECAT and CCprofiler outperformed EPIC in all pairwise comparisons, especially in the high-confidence region, potentially because that algorithm was not primarily developed for the analysis of SEC, but rather ion-exchange (IEX) profiles [31] (Supplementary Fig. 2b-d).

In summary, the PPI detection module provides accurate estimation of PPI detection confidence which is achieved by combining a targeted approach with a semi-supervised learning strategy. Differential quantification of network states requires

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

both PPI detection reproducibility and sensitivity. Compared to algorithms that require peak picking, SECAT's minimal signal preprocessing supports more reproducible PPI detection across samples, while still performing competitively on single replicates. While SECAT can be applied to data from other PCP sample preparation and data acquisition methods, it benefits from the quantitative accuracy of DIA-based SWATH-MS.

Validation of PPI quantification & network integration modules

Building on SECAT's benchmarked ability to detect true PPIs from PCP datasets, we next validated its PPI quantification and network integration modules. For these analyses, we used the previously described HeLa-CC dataset [38]. For this study, we performed independent analyses, using CORUM [37], STRING [19] and PrePPI [24] as reference PPI networks, and used a classifier trained on the full CORUM true/false reference network.

PPI detection with different reference networks: For the three reference PPI networks, the total query spaces ranged from 15,939 to 449,265 PPIs (Fig. 3a). Of these, between 5,460 and 9,026 high-confidence PPIs (q-value $< 5\%$) were detected, with a core set of 3,306 common PPIs (Fig. 3b). Notably, the full intersection of detected true PPIs might be larger, but two factors conspire to make objective comparison more complex. First, different confidence scores provided by each database are used as priors in our analyses. Second, the largely variable number of PPIs represented in each database needs to be accounted for during multiple testing correction (Methods).

PPI quantification and relation of metric classes: SECAT's quantification module computes two main quantitative property categories from SEC-SWATH data: one indicating protein abundance, the other representing PPIs. Some of these metrics are expected to be related. For example, if a protein is only present in an "assembled" conformation, the "total abundance" and "assembled abundance" values will be highly correlated. To assess the relation and redundancy within metric classes, we conducted dimensionality reduction using principal component analysis (PCA), separating replicates and conditions (Fig. 3c). The data show that the first two principal components of the "total abundance", "monomer abundance" and "assembled abundance" metrics explain 60.72-66.48% of the variance between them, while the first two principal components for the "interactor abundance", "complex abundance" and "interactor ratio" metrics explain 62.66-73.93% of the variance. The small difference in variance between protein abundance- and PPI-based metrics might potentially arise from metrics quantifying redundant interactions between same-complex proteins.

To assess the relationship between metric classes, we further computed the distribution of Spearman's correlation of metrics aggregated over conditions, replicates and proteins to the corresponding "total" protein abundances (Fig. 3d). As expected, the "assembled abundance" metrics which cover the majority of SEC fractions have the strongest correlation, followed by "monomer abundance", "interactor abundance" and "complex abundance" metrics. Notably, since the "interactor ratio" metrics summarize relative changes between the interactors, they are substantially less correlated.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

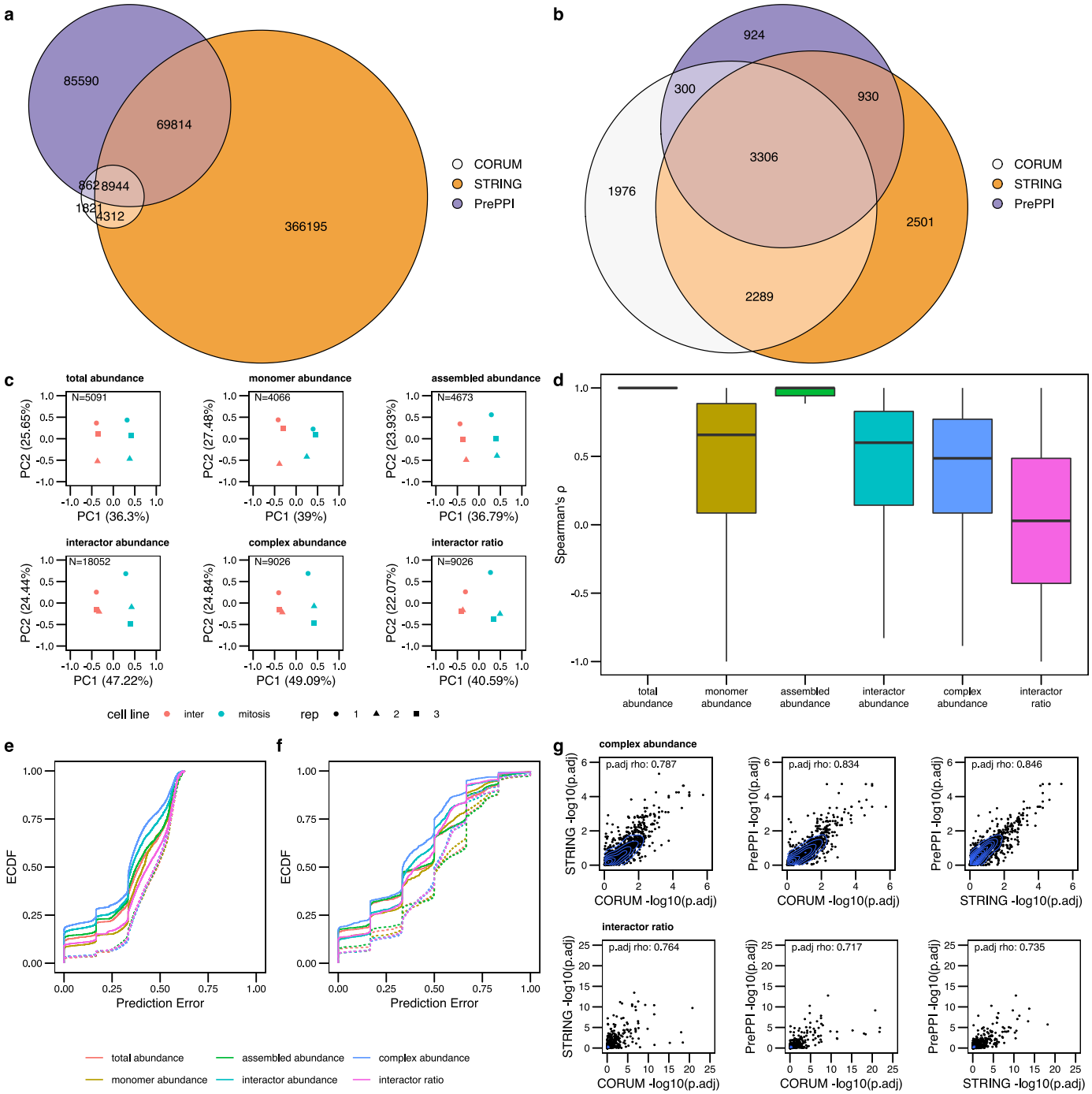


Figure 3. PPI quantification and Network integration assessment using the HeLa-CC [38] dataset. (**a**) The Euler diagram indicates the overlap of STRING, PrePPI and CORUM reference PPI networks resulting in at least a partial overlap of the interactors in any of the replicates. (**b**) Overlap of confident (q-value < 0.05) PPI detections after SECAT analysis with different reference PPI networks. (**c**) Quantitative SECAT metrics can be used to separate experimental conditions and replicates. (**d**) Spearman's correlation of the metrics (PPI: summarized by averaging) to total-level is depicted in individual boxplots (lower and upper hinges represent the first and third quartiles; the bar represents the median; lower and upper whisker extend to 1.5 * IQR from the hinge). (**e**) Predictive error quantified by the empirical cumulative distribution function (ECDF) for leave-one-out cross-validated logistic regression classification of

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

replicates to biological conditions. (**f**) Predictive error for group-wise leave-one-out cross-validated logistic regression classification of replicates to biological conditions. Groups represent several metrics connected to a core metric based on the PPI network. (**g**) Correlation of node-level integrated adjusted p-values estimated using different reference PPI networks.

Assessment of predictive power of quantitative metrics: To assess the predictive power of the different metric classes for classifying the six samples to one of the two represented mitotic states, we estimated the prediction error using leave-one-out cross-validation using a logistic regression classifier applied to each metric separately (Fig. 3e). The empirical cumulative distribution function (ECDF) in dependency of the prediction error indicates that the "complex abundance" metrics are most informative, followed by "interactor abundance" and the other metrics, to separate the two mitotic states. While these metrics are computed for single proteins or PPIs, network-centric data integration might further boost the predictive power of the respective values.

To assess this effect, we combined individual PPI metrics based on the SECAT PPI networks by extending the score vector for the classifier from single metrics to all PPI scores connected to a specific protein (Fig. 3f). The results show that the "complex abundance"-based metrics performed best and the "assembled abundance" metrics performed second best, indicating that the fractions covering assembled protein conformations have a higher information content than those covering monomeric subunits and that the connectivity information calculated by SECAT has higher information content than the protein abundance.

Redundancy of network integration: In higher order complexes composed of three or more subunits, PPI information is expected to be partly redundant, with each subunit relating to at least two other co-complex members. Thus, a change in one subunit might be apparent in other PPIs as well. This PPI redundancy between subunits of a multiprotein complex is critically relevant to any strategy attempting to integrate individual protein or PPI properties at the network level. Consistent with this observation, SECAT assumes that changes in the individual subunits of a protein complex can be measured redundantly by assessing its interactions with all proteins in the complex. Different reference PPI networks are thus expected to provide comparable quantitative metrics if they partially overlap. To test this assumption, we compared the differential integrated metrics on "complex abundance" and "interactor ratio" levels using different reference PPI networks. The results indicate a high degree of correlation among the "complex abundance" (Spearman's rho: 0.787-0.846) and "interactor ratio" levels (Spearman's rho: 0.717-0.764) (Fig. 3g), confirming that different reference PPI networks provided comparable quantitative metrics from the same PCP input data.

In summary, SECAT's PPI quantification and network integration modules generate specific PPI networks for each sample or phenotypic state. To achieve this, the signal processing and PPI detection modules first transform SEC fraction peptide abundance values to a set of protein abundance or complex-associated metrics. The presented validation results show that PPI-level quantitative metrics have higher predictive power to separate sample groups than the underlying total protein abundances. In a second step, redundant and non-redundant PPI-level metrics (edges) are integrated to protein-level (nodes). With different empirical or predicted reference PPI networks, this transformation achieves similar results attesting to the robustness of results despite different background PPI networks.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Molecular mechanisms differentiating HeLa cell cycle states

To demonstrate SECAT's ability to discriminate molecular mechanisms that are differentially represented across experimental conditions or phenotypes, we further investigated the results obtained from the HeLa-CC dataset [38]. This study was designed to compare the protein complex differences between cell cycle states of a HeLa CCL2 cell line, first, by thymidine blocking (arrest in interphase) and then second, by release in nocodazole (arrest in mitosis) [38]. For each experimental condition, three full-process replicates were generated, and 65 SEC fractions per condition and replicates were measured by SEC-SWATH-MS. Cumulatively, 70,445 peptides associated with 5,514 proteins were quantified across the full dataset. We conducted the SECAT analysis of this quantitative matrix using STRING as the reference PPI network with the goal to identify phenotype-associated molecular mechanisms (Methods). Since the cell cycle and its checkpoints are controlled by clearly defined events involving selected checkpoint proteins, this provides a reference framework for the interpretation of the results.

After network-centric data integration, SECAT identified 28 differentially abundant proteins ("total abundance"-level) across the two cell states, 50 alterations affecting "assembled abundance", no alterations regarding "monomer abundance", 110 alterations affecting "interactor abundance", 117 alterations in "complex abundance", and 139 alterations in "interactor ratio" (adj. p-value < 0.01; $|log2(\text{fold-change})| > 1$) (Supplementary Data 2-3). Since, as discussed above, different data aggregation levels can be highly correlated, SECAT implements for visualization purposes a simplistic aggregation to the most significant level per protein that provides a reduced overview. This can be further augmented by integration of the context-specific PPI network with Reactome [49] (Fig. 4, Methods).

GSEA analysis of the top-level Reactome pathways (Fig. 4a) shows that protein-complex rearrangements constitute the primary significant differences between the two cell states. Proteins associated with the categories "Cell Cycle", "Metabolism of RNA", "DNA Replication" and "Protein localization" were especially affected. Two primary clusters detected by the analysis, comprising 88 and 74 proteins, respectively, were associated with translation and mitochondrial translation respectively (Fig. 4b). Further, large modules that differ between cell states include "respiratory electron transport" (72 proteins), ribosomal RNA processing (57 proteins) and RNA metabolism (36 proteins).

Cyclin B1 bound to Cyclin-dependent kinase 1 (Cdk1) is a major catalytic factor promoting mitosis [50]. Counteracting Cyclin B1-Cdk1 mediated activation, sequential degradation of cell cycle regulating proteins via the ubiquitin pathway is important to progress through mitosis [51]. SECAT recalled these well-known biochemical events from the SEC-SWATH data. Specifically, it detected different levels of Cdk1 "complex abundance" and "interactor ratio" as a principal factor differentiating the two cell cycle states (Fig. 5a). Further, SECAT identified several subunits of the anaphase-promoting complex (APC) to be significantly more abundant during mitosis (Fig. 5b). The network visualization further illustrates the high connectivity between the individual subunits, resulting in a distinctive complex module within the graph, primarily affected by differential protein abundance between the mitosis states (Fig. 4b). This is consistent with the role of the APC as a ubiquitin ligase mediating this particular step [51].

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

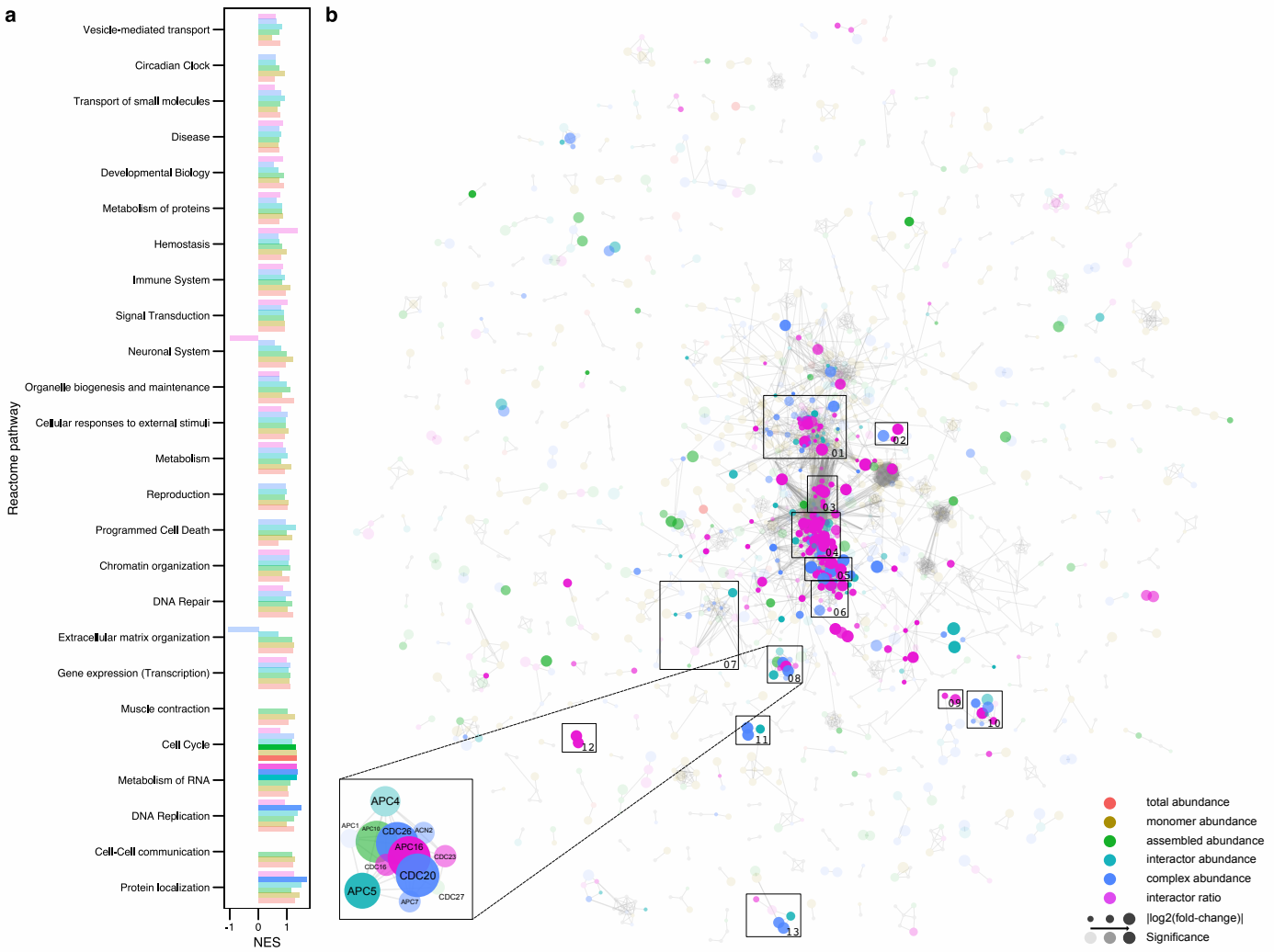


Figure 4. Molecular mechanisms differentiating HeLa cell cycle states. (**a**) Gene set enrichment analysis indicates that changes on protein complex abundance-level are dominant. Opaque categories are significant (adj. p-value < 0.05). (**b**) The integrated STRING-based PPI network of the HeLa cell cycle dataset [38] is depicted with proteins (nodes) and binary PPI (edges) clustered against Reactome pathway association (Methods). Different colors indicate the most significant metric-level (legend, Methods). Node-size indicates effect-size and opacity indicates significance. The most prominent clusters highlight essential macromolecular assemblies: (1) Respiratory electron transport, (2) Mitochondrial protein import, (3) Translation, (4) rRNA processing, (5) Metabolism of RNA, (6) mRNA slicing, (7) Nuclear pore complex proteins, (8) Anaphase-promoting complex (APC), (9) Activation of the TFAP2 (AP-2) family of transcription factors, (10) SWI/SNF complex, (11) Condensin-2 complex, (12) Chromatin modifying enzymes, (13) Origin recognition complex. The inset shows the APC subunits and highlights their connectivity and changes on different subunit levels. Interactive results (Cytoscape) are provided as Supplementary Data 1.

Between transcription and translation, pre-mRNA is processed to remove introns and splice exons to produce mature mRNA molecules for translation. This process is catalyzed by the spliceosome, a multi-megadalton ribonucleoprotein complex, which highly dynamically adapts to context-dependent functions [52]. Spliceosome assembly and function typically involve several intermediate complexes, requiring the integrity of the nuclear compartement [40]. With the disassembly of the nucleus and associated nuclear pore complex (NPC) proteins during mitosis, the rate of transcription is reduced and it is currently believed that spliceosome subunits are distributed across the

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mitotic cytoplasm awaiting re-activation upon nuclear reassembly [40]. However, systematic screens identified spliceosome subcomplexes, including the NineTeen Complex (NTC) with five out of seven of its core proteins (PRPF19 (PRP19), CDC5L, SPF27, PLRG1 and CTNNBL1 (CTBL1)) as essential components for mitosis [53]. Correspondingly, in our analyses PRPF19 and CDC5L were identified as differentially abundant between interphase and mitosis on the "complex abundance" level. In addition, they display changes on the "interactor ratio" level. Specifically, the NTC subunits form a more distinctive SEC elution peak during mitosis (Fig. 5c). Similarly, NHP2-like protein 1 (NH2L1), a component of the U4/U6.U5 tri-snRNP subcomplex, has been found to be required during mitosis [53]. Our analysis further supports the importance of NH2L1 during mitosis, as it is among the most significantly changed proteins in terms of "interactor ratio" but not "total abundance". SECAT further found the linked U4/U6.U5 tri-snRNP subcomplex proteins such as the tri-snRNP-associated protein 2 (SNUT2), the pre-mRNA-processing factors 3, 4, 6 and 31 (PRP3,4,6,31) to be of similar differential "interactor ratio" significance, suggesting subcomplex activity during the cell cycle (Fig. 5d).

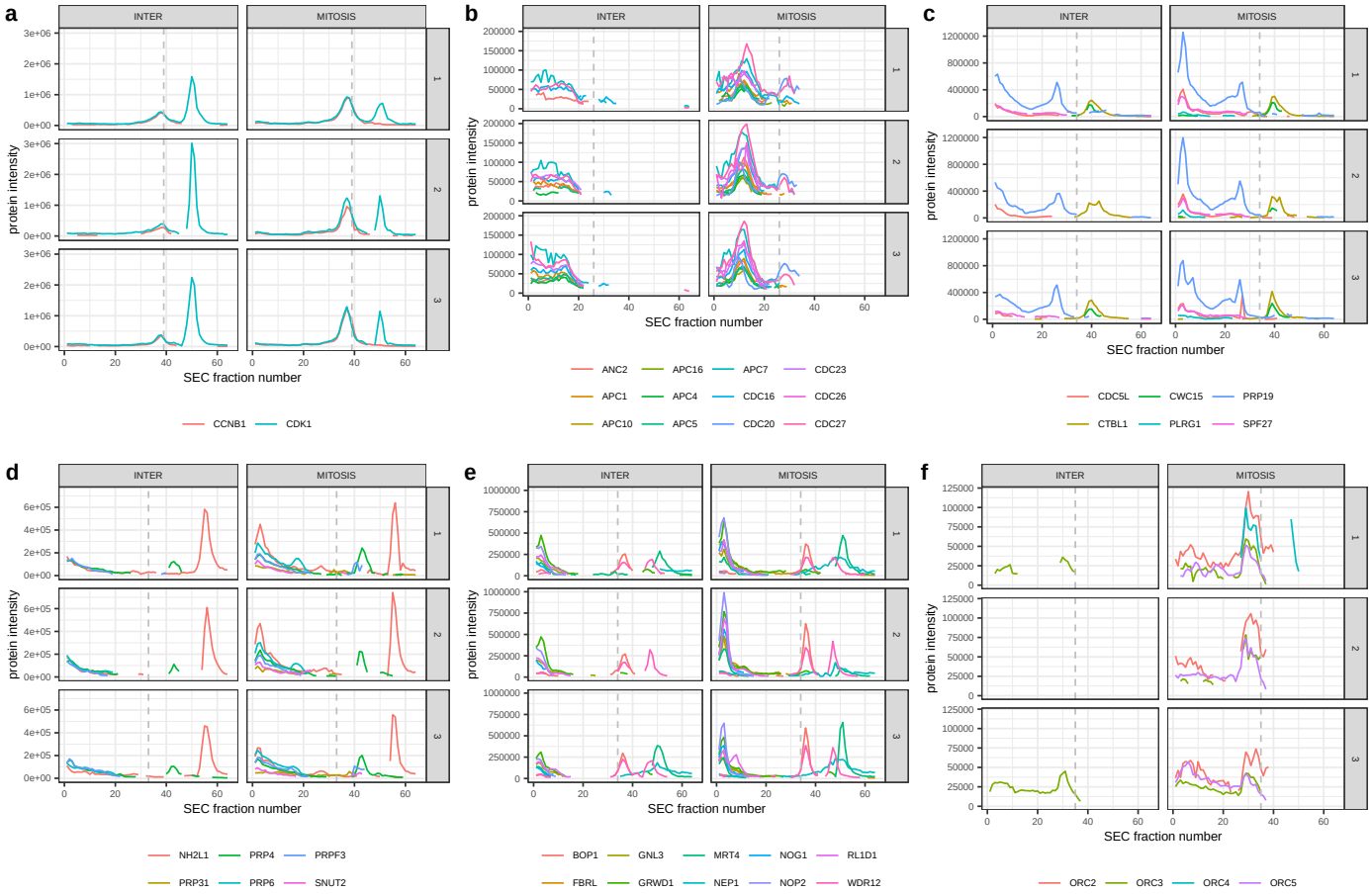


Figure 5. Protein-level SEC-SWATH profiles of HeLa cell cycle states. Lines indicate different protein subunits, whereas the dashed line indicates the highest monomer threshold per group. (**a**) Cyclin B1-Cdk1 complex. (**b**) Subunits of the anaphase-promoting complex (APC). (**c**) Subunits of the NineTeen Complex (NTC). (**d**) Associated U4/U6.U5 tri-snRNP subcomplex proteins. (**e**) The ten most abundant ribosome biogenesis-associated proteins. (**f**) Subunits of the origin recognition complex (ORC).

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Ribosomal RNA Processing complex represents one of the larger and most densely connected submodules of the dataset. SECAT found 40 out of total 57 proteins known to belong to this group to be differentially abundant (Fig. 5e). Of those, the majority is associated with ribosome biogenesis, a process located in both cytoplasm and nucleolus. Similar to the origin recognition complex (ORC, Fig. 5f), which is only active in the nucleus, this difference in abundance might be explained by the experimental design, because under the conditions used for cell lysis, proteins of the nucleus are less accessible during interphase than mitosis, a cell cycle state where the NPC is disassembled.

In summary, SECAT provides context-specific networks and protein-level metrics that can be visualized as intuitive maps (Fig. 4b), facilitating the interpretation of observed molecular differences between cell states at different levels including protein abundance, PPIs, protein complexes and PPI network modules, concurrently, from the same dataset.

Discussion

Protein-protein interactions within protein complexes are a principal characteristic of the proteome and significantly affect the biochemical state of the cell. Interaction proteomics has developed powerful methods such as AP-MS to investigate the interactions of specific proteins at relatively high throughput. The cumulative results of thousands of AP-MS measurements constitute PPI network maps for investigated organisms [16, 18]. However, the extension of the AP-MS approach to compare PPI networks at different states is intrinsically limited because it would require the comparative analysis of the results of a high number of AP-MS measurements in the different states.

For this reason, PCP-based methods, such as SEC-SWATH-MS, have emerged as complementary approaches. They can measure protein complex profiles for thousands of proteins, quickly and reproducibly. They achieve, however, lower proteome coverage and are limited to medium to high affinity-binding protein complexes that are soluble and remain intact under the extraction conditions used. Previous studies have already demonstrated the application of PCP to qualitatively characterize metazoan macromolecular complexes [29]. With increasing throughput and further methodological improvements, it can be expected that these developments enable the qualitative and quantitative comparison of dozen to hundreds of samples in single studies.

However, the relatively low peak capacity of SEC imposes major limitations for PPI identification, i.e. the number of proteins identified by far exceeds the number of separable peaks and thus fractions collected. In previous studies this limitation was addressed by sequentially applying orthogonal biochemical fractionation methods [27] or by focusing on protein complex detection using predetermined subunits [3]. This requires either more complex and costly experimental designs or focus on well characterized protein complexes, limiting the scalability and generic applicability of protein complex profiling studies.

With SECAT, we propose an alternative analysis strategy, which makes use of the high consistency of peptide-level quantification of SWATH-MS and prior knowledge from PPI reference databases. We demonstrate that SECAT applied to data with these qualities provides accurate estimation of PPI detection confidence while substantially increasing the coverage of binary interactions. The robustness of the scoring and

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

semi-supervised learning strategy further permits omission of preprocessing steps such as peak picking, making the algorithm robust against context-specific deviations related to SEC peak shape or different calibration of SEC fractions.

Because SEC separates stable native protein complexes, the inference of their composition from binary interactions is a key component of most previous data analysis strategies. This provides the opportunity to identify previously unknown associations in a global fashion, however, the underlying challenges of annotating and comparing context-specific related subcomplexes will become more severe with increasing number of conditions tested in a study. We show that the proposed protein abundance-level and PPI-based metrics are comparable across different PPI reference networks and thus their implementation in SECAT provides a scalable alternative to protein complex inference that makes use of redundant information to increase the consistency of quantitative comparisons.

Our application of SECAT to the HeLa-CC dataset [38] illustrates that different network states can efficiently be visualized in a network-centric representation to highlight the complex relations of different qualities. This provides a bird's-eye view of alterations in PPI networks that can be used intuitively to guide follow-up investigations. SECAT is available for all platforms as open source software implemented in Python and is compatible with different LC-MS/MS profile and reference PPI database formats. We expect that our toolkit and the underlying concepts will be particularly useful for future PCP datasets, guiding the qualitative and quantitative comparison of multiple conditions, where protein complexes represent dynamic rather than static modules.

Methods

The Size-Exclusion Chromatography Algorithmic Toolkit

Peak picking

Protein complexes are expected to form Gaussian-like elution peaks in SEC. While this assumption holds true for some protein subunits, the high dynamic range (3-4 orders of magnitude) of protein abundance measured by LC-MS/MS in combination with varying peak width and convoluted (sub-)complexes result in a diverse set of elution peaks. While increasing sensitivity and quantitative consistency of DIA-based methods like SWATH-MS provide more accurate peptide and protein profiles, the fundamental limitations, particularly the low resolution and the logarithmic-linear relation between the apparent molecular weight and the fraction index of SEC pose great analytical challenges.

To boost signal-to-noise ratio (SNR), PrinCE [34] and CCprofiler [3], implement peak-picking via Gaussian deconvolution or sliding window peak detection respectively. The resulting sets of candidate peaks for each protein are more uniform and co-eluting peaks can be scored and grouped to the same protein complex. The downside is that non-Gaussian peaks or highly convoluted peaks are more difficult to detect. Other approaches [29] compute full elution profile correlation profiles between potential interacting proteins, which has the advantage that peak picking is not limiting, but the downside that proteins that are present in different complex conformations will potentially not be detected.

SECAT adapts a third approach, where only the intersection of the queried proteins is assessed with the following assumptions: First, two candidate interactors are expected to interact via the same molecular mechanism as part of their constitution in different

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

protein subcomplexes. Second, if this interaction is changed or perturbed via a specific mechanism, the subcomplexes will be similarly affected. The benefit of this approach is that the involvement of the protein subunits as part of different, non-overlapping complexes or monomers can be independently assessed. Further, the omission of peak-picking enables consistent scoring and quantification across multiple samples, a key property when heterogeneous systems or samples acquired with different SEC configurations need to be compared.

Optionally, SECAT provides two peak-picking methods for preprocessing data before scoring. "Detrending" subtracts the mean of the peptide abundance over all SEC fractions, where missing values are replaced by zero. All fractions with negative values are then excluded and the original intensities for the other fractions are retained. The "localmax" peak picking algorithm computes the mean of all peptide intensities per protein and requires a minimum SEC fraction number of 3 continuous fractions, while setting peak boundaries at a relative height of 0.9. Only peptide intensities of SEC fractions within peaks are reported.

Peak scoring

Several scoring systems for SEC and related separation techniques have been developed. They have in common that the inferred protein-level abundance profiles over the SEC dimension are used for scoring. The inferred protein-level abundance provides a robust estimate, especially when peptide quantification is less sensitive or quantitatively consistent, as is the case with older LC-MS/MS setups or DDA-based data acquisition.

In contrast, our recently developed technique SEC-SWATH-MS [3] can provide accurate peptide quantification across the full SEC fraction range. Because CCprofiler in complex-centric mode assesses two or more individual subunits, peptide abundances are aggregated before peak picking and scoring.

To optimally use the additional data points that can be assessed by comparing consistent peptide quantification profiles, SECAT uses variable numbers of peptide traces for scoring (default: 1-3). Inspired by our previous scoring systems for chromatographic data [42, 43, 54], cross-correlation shape and SEC fraction shift across two candidate interacting proteins are assessed. SECAT further estimates the maximal and total information coefficients (MIC/TIC) [44] to assess the statistical association between the overlapping sections of the SEC profiles of the two proteins. Protein complexes and their binary representations follow specific rules in terms of interactor ratio. To capture this property, we estimate the abundance ratio between the individual subunits. Finally, the delta number of fractions between the apex of the dominating candidate PPI peak and the monomer threshold is computed.

Classification by semi-supervised machine learning

While the partial scores describe different properties that can be used to separate true from false candidate PPI, they are ideally integrated to a single discriminant score via classification [34]. Further, establishing a null model that represents spontaneous co-elution of individual protein subunits is required to estimate the confidence of PPI detection accurately.

SECAT adapts a semi-supervised learning strategy inspired by the approaches implemented in Percolator [55], mProphet [42] and PyProphet [45, 54]. As positive reference dataset, SECAT uses the PPI covered by the CORUM [37] reference database. However, in contrast to the strategies employed for scoring spectra or fragment ion chromatograms, the SEC profiles of potential interactors can be more heterogeneous, and a single score might not be discriminative enough to select the best positive

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

examples for semi-supervised learning. For this reason, SECAT computes a "kickstart" score, that integrates cross-correlation shape, co-elution and mass ratio to separate the very best scoring examples from CORUM during the first iteration. Semi-supervised learning is a crucial component of the SECAT scoring model, because the omission of peak picking generates random, partially overlapping protein profiles that would be used as true interactions in the classifier otherwise.

Different approaches have been described to generate null models for PPI identified or detected using SEC data. Random permutation of PPI can potentially generate large sets of negative interactions, but the later restriction of the set to at least partially overlapping SEC profiles might potentially increase the fraction of true interactions. Ground truth negative interactome reference databases like IntAct [20] or Y2H-predicted [56] pose an alternative, but are either low in coverage or do not represent the target reference databases well. By default, SECAT thus uses the CORUM-inverted reference database as was previously proposed [34]. The benefit is that the target protein set is well represented, and a large number of negative PPI can be obtained. To minimize the number of potential true PPI not covered by CORUM, all overlapping PPI with reference databases are further removed.

SECAT uses XGBoost [57], a tree boosting approach to maximize classification performance. XGBoost replaces the linear discriminant analysis (LDA)-based classifier used in PyProphet during all learning iterations with hyperparameters that can optionally be automatically tuned during semi-supervised learning. In a first step, SECAT uses the positive and negative reference databases to learn a classifier using the kickstart score. At each subsequent iteration (default: 10), an XGBoost [57]-based classifier is then trained on the true positives, with an initial q-value of 0.1 and iterative q-value of 0.05, from the prior iteration—starting from the initial score based ones—and used to generate increasingly discriminative integrative scores from individual partial scores. By default, three cross-validation iterations are conducted. "Early stopping" is enabled in the XGBoost learning step to alleviate overfitting. The scores of the final learning iterations of each cross-validation step are then averaged and used to train a final classifier. In a second step, statistical validation is conducted using the Storey-Tibshirani framework [58]. At this step, prior information about PPI confidence from reference databases can be incorporated: SECAT generates N (default: 100) quantiles for the PPI confidence scores and applies statistical validation separately. This enables multiple hypothesis testing correction to account for different prior probabilities of false detection of interactions and is generically applicable to confidence scores with different statistical properties.

Quantification of protein-protein interactions

SEC peptide profiles can be partitioned into components to represent quantitative information on different levels. The sum of the full elution profile corresponds to the total peptide abundance, which to some degree represent protein abundances measured by conventional, non-fractionated LC-MS/MS. In a first step, SECAT extracts and adds up the monomeric components by defining a threshold fraction for each protein profile based on the estimated monomeric weight multiplied by a constant factor (default: 2) to account for uncertainty in the apparent molecular weights of the SEC fractions. Complementarily, the fractions on the left hand-side of the monomer threshold are summarized and represent the assembled protein conformations. To compute the PPI-level quantitative metrics, for each PPI below a specified confidence threshold (default: q-value < 0.05 in any of the compared conditions), the intersecting fractions of the two interactor protein profiles are analogously extracted on peptide level.

Quantitative protein inference is conducted using proteoVIPER, which is based on

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the VIPER [41] algorithm. Originally developed for inference of protein activity using transcriptomics data, instead of a gene regulatory network, the peptide-protein relationships are used as for the analytical rank-based enrichment analysis (aREA) to infer the scores for the protein abundance-based metrics. For the PPI-based metrics, the model is slightly adjusted: To assess protein complex abundance, the peptides for both interactor proteins are used with the same mode of regulation. For scoring protein complex interactor ratio on the other hand, the mode of regulation has different signs for each of the interactors, which enables proteoVIPER to assess changes in interactor ratio. The implementation of proteoVIPER in SECAT infers a quantitative matrix with normalized enrichment scores (NES) for each condition and replicate. This approach has the benefit that proteins with missing values or not perfectly overlapping peptides can be accurately and robustly quantified. Experimental conditions, from prior annotation or selected based on the data, can then be statistically compared by independent t-tests [41] on each level.

Network-centric data integration

Using the quantitative metrics from above, SECAT conducts network-centric data integration. The concept is based on the assumption that specific experimental conditions (e.g. drug perturbations) have a specific effect on individual protein subunits that can be detected via one or several of their interactions.

The evidence of multiple measured PPI is summarized to protein complex metrics summarizing changes in protein complex abundance or interactor ratio for each protein using Empirical Brown's Method [47]. Notably, highly correlated PPI (e.g. from the same protein complex) are integrated in a dependent fashion, whereas independent PPI (e.g. from different protein complexes) combine and increase the significance of the protein complex engagement metric.

In a network context, this helps to identify the most perturbed or dysregulated proteins based on changes of their protein complexes. Instead of clustering or inferring protein complexes, which are difficult to define in presence of subcomplexes across multiple experimental conditions, SECAT's metrics can be more robustly characterized from a partial subset of their interactions.

Integrated p-values are adjusted for multiple testing using the Benjamini-Hochberg [59] approach, as suggested [47].

Primary data analysis

Processed mass spectrometry data has been obtained from the repositories linked by the original publications or the authors of the corresponding publications.

The quantitative values on peptide level were normalized via cyclic lowess [60, 61] grouped by a sliding window of 5 units and a step-size of 1 over the SEC fraction indices before processing in SECAT (Supplementary Fig. 3-6).

SECAT data analysis

SECAT (version 1.0) and PyProphet (version 2.1.0) were used for all data analyses with CORUM (version 3.0 [37]), PrePPI (version 2016 [24]) and STRING (version 11.0 [19]) and default parameters if not otherwise specified. Semi-supervised learning was conducted using CORUM as positive network and CORUM-inverted as negative network. All input data and parameters are provided on the Zenodo repository.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CORUM-inverted was generated by using the inverted set of PPI (i.e. all possible PPI that are not covered by CORUM) and removing all PPI in this set covered by STRING [19], IID [62], PrePPI or BioPlex [18].

Parameter selection and validation of signal processing and PPI detection modules

The SECAT PPI detection benchmark was conducted by using 50% of the CORUM reference PPI network for learning and the other fraction for evaluation. Reference false PPI from CORUM-inverted were randomly selected in predefined ratios (1:0 – 1:16) and added to the target set for evaluation but not learning.

Fig. 2a and Supplementary Fig. 1a depict violin plots with the following parameters: Lower and upper hinges represent the first and third quartiles; the bar represents the median. This represents the default parameters of the function "geom_violin" of ggplot2.

Fig. 2b and Supplementary Fig. 1b were generated by assessing the PPIs with a global-context q-value < 0.05 and decomposing the number of PPIs for detection amongst replicates at different confidence thresholds.

Fig. 2e and Supplementary Fig. 1e were generated by using the ground truth CORUM and CORUM-inverted reference values. Because the estimated q-values are dependent on the combined reference sets with unknown ratios of true and false PPIs, the "true q-values" were corrected by the estimated proportion of true positives by PyProphet.

For Supplementary Fig. 2, the CORUM reference PPI network was similarly used as described above to generate a pseudo-ground truth dataset for classifier training. However, for the validation subset, an excess of 10 times as many false interactions as targets prior to analysis was added. For the downstream analysis, the CORUM targets only were reduced to the intersection with a manually curated annotation of the dataset [3].

The CCprofiler [3] analysis (git revision: efdeca4) was conducted as suggested by the software documentation. All input data and parameters are provided on the Zenodo repository.

The EPIC [31] analysis (git revision: b6432b9) was conducted as suggested by the software documentation with the provided Docker container. The input data was first aggregated from peptide-level to protein-level using the top3 method implemented in aLFQ [63] (version 1.3.5). All input data and parameters are provided on the Zenodo repository.

Validation of PPI quantification & network integration modules

The data was analyzed as described above with the full CORUM, PrePPI and STRING reference networks. Fig. 3c-f were generated using the STRING-based analysis. Fig. 3d depicts boxplots with the following parameters: Lower and upper hinges represent the first and third quartiles; the bar represents the median. Lower and upper whisker extend to 1.5 * IQR from the hinge. This represents the default parameters of the function "geom_boxplot" of ggplot2.

Molecular mechanisms differentiating HeLa cell cycle states

The gene set enrichment analysis in Fig. 4a was generated using the R-package "fgsea" [64], applied on each level separately. Data was prefiltered ($|log2(\text{fold-change})| > 1$ or differential interactor ratio < 0.9) on each level separately

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and -log10(BH-adjusted p-value) was used as score. The Reactome Pathway Database [49] (version 67; only top-levels) was used with 10,000 permutations and a minimum gene set size of 5; all proteins of the SEC-SWATH-MS dataset were used as background. Opacity in the figure indicates significance (BH-adjusted p-value < 0.05).

To annotate and visualize differential proteins in Fig. 4b between the HeLa cell cycle states identified by SECAT, we used Cytoscape [65] (version 3.7.1). The Reactome Pathway Database (version 67) was used to cluster PPI using the Cytoscape App AutoAnnotate [66] with default parameters and a maximum cluster size ("Max words per label") of 1. Clusters were arranged according to the CoSE layout.

Visualization of protein-level SEC-SWATH-MS profiles in Fig. 5 was conducted using the R-package "ggplot2" by averaging the three most intense peptide precursors per protein.

Source code availability

SECAT is available as platform-independent open source software under the Modified BSD License and distributed as part of the SECAT (`https://pypi.org/project/secat`) and PyProphet (`https://pypi.org/project/pyprophet`) Python PyPI packages. SECAT further depends on the R/Bioconductor package "viper", which is distributed under a non-commercial usage license. Further documentation and instructions for usage can be found on the SECAT source code repository (`https://github.com/grosenberger/secat`).

Data availability

All analysis results are available on Zenodo with the dataset identifier 10.5281/zenodo.3515928.

Acknowledgments

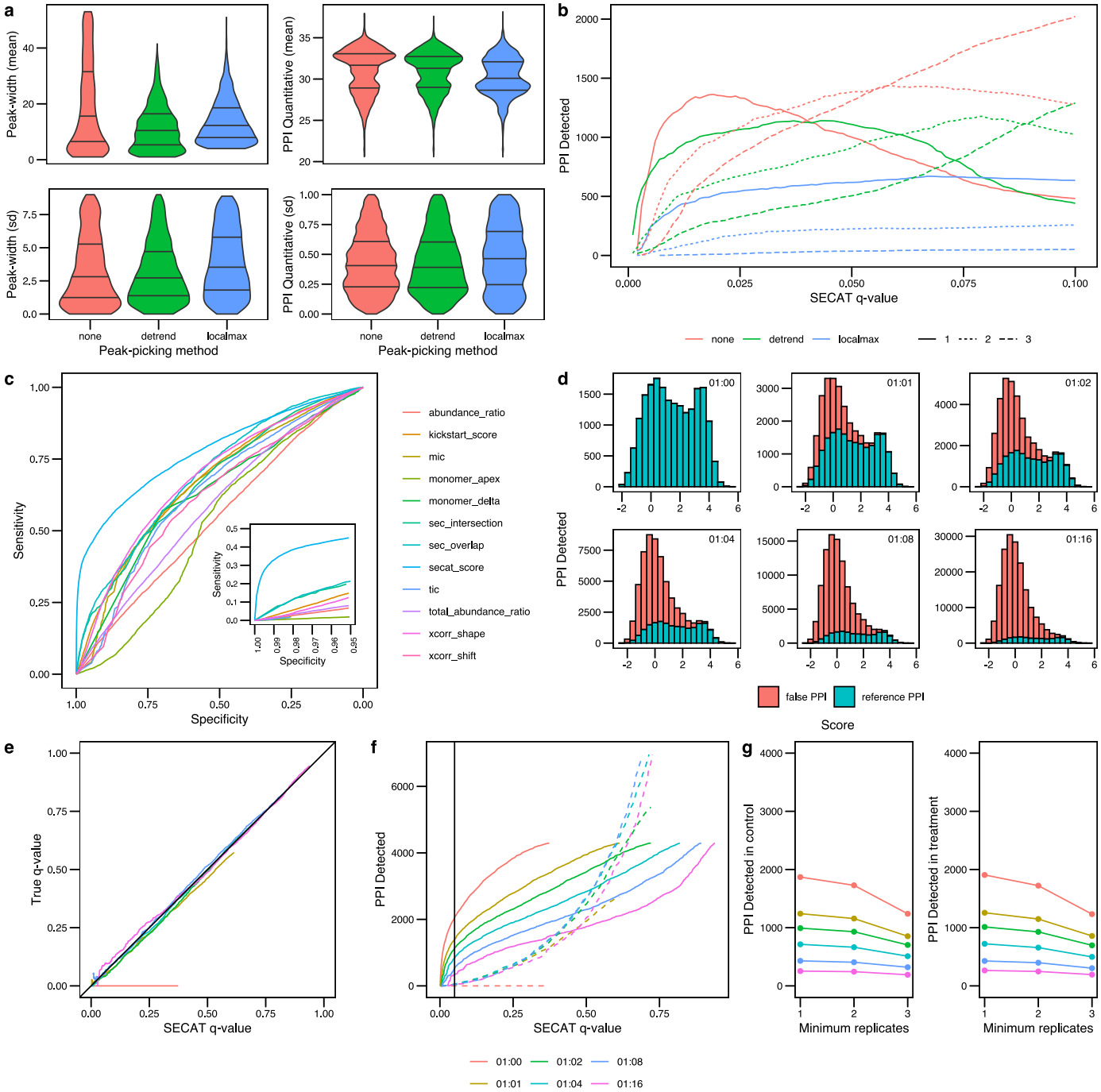
We thank N.E. Scott and L.J. Foster for providing access to the peptide-level data of their study [30]. G.R. is supported by grants P2EZP3_175127 and P400PB_183933 from the Swiss National Science Foundation. The project was supported by the European Research Council (ERC-20140AdG 670821) and the Swiss National Science Foundation (grant 31003A_166435 to R.A.).

Author contributions

Conceptualization, G.R., M.H., R.A. and A.C; Formal analysis, G.R.; Funding acquisition: G.R., R.A.; Investigation: all authors; Methodology: G.R., M.H., I.B.; Resources: R.A. and A.C.; Software: G.R.; Supervision: R.A. and A.C.; Validation: G.R.; Visualization: G.R.; Writing – original draft: G.R., R.A. and A.C.; Writing – review & editing: all authors.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figures

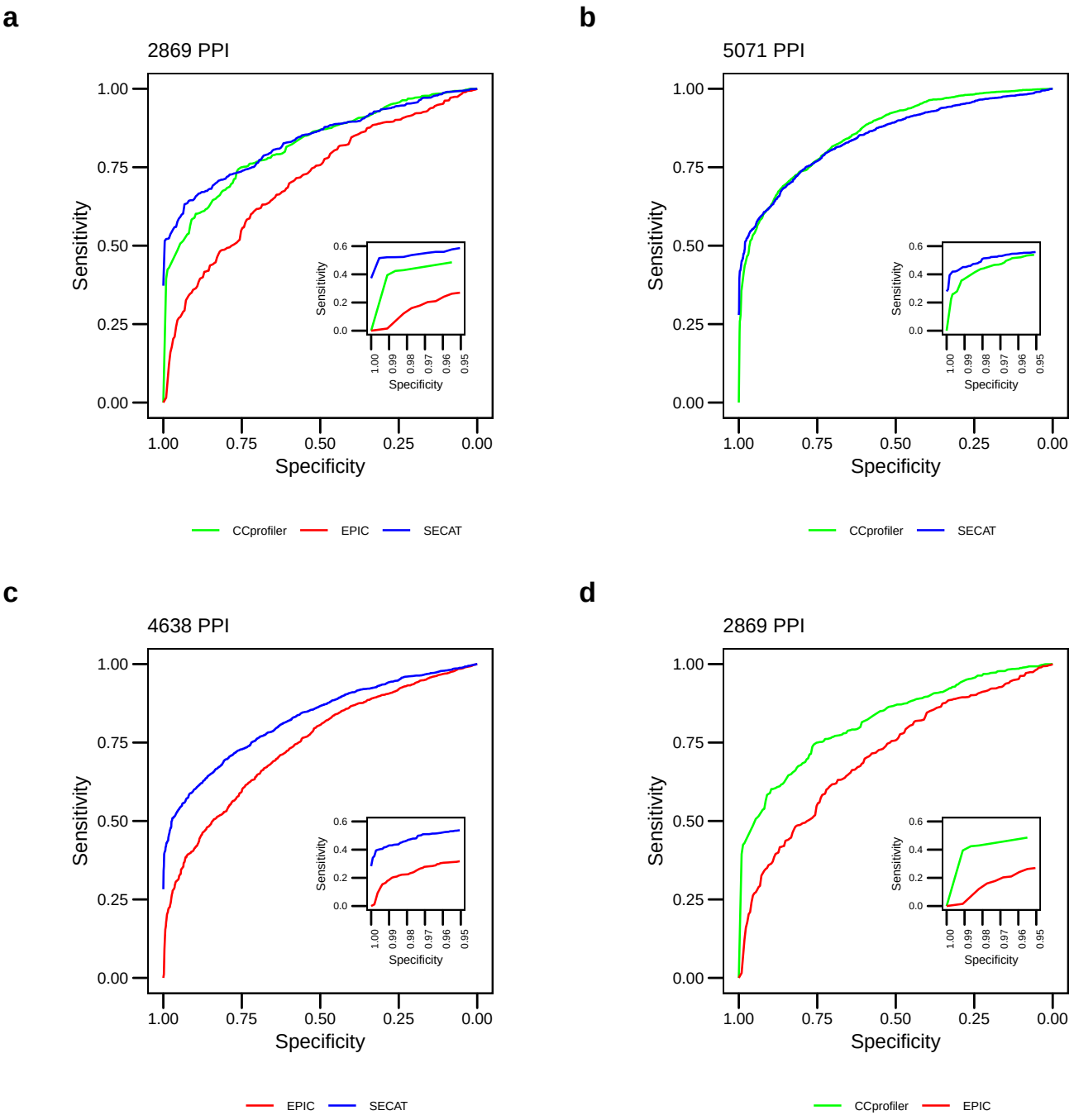


Supplementary Figure 1. Signal processing and PPI detection performance evaluated using the Jurkat-Fas [30] dataset. (**a**) The effect of different peak-picking methods for signal processing on peak-width and PPI quantification within replicates of the same experimental condition is depicted in violin plots (lines representing 25, 50 and 75% quantiles respectively, Methods). (**b**) Sensitivity of PPI detection vs. SECAT q-value for peak-picking options. The PPI were filtered to a global context q-value < 0.05. Colors indicate different peak-picking options; line types indicate

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

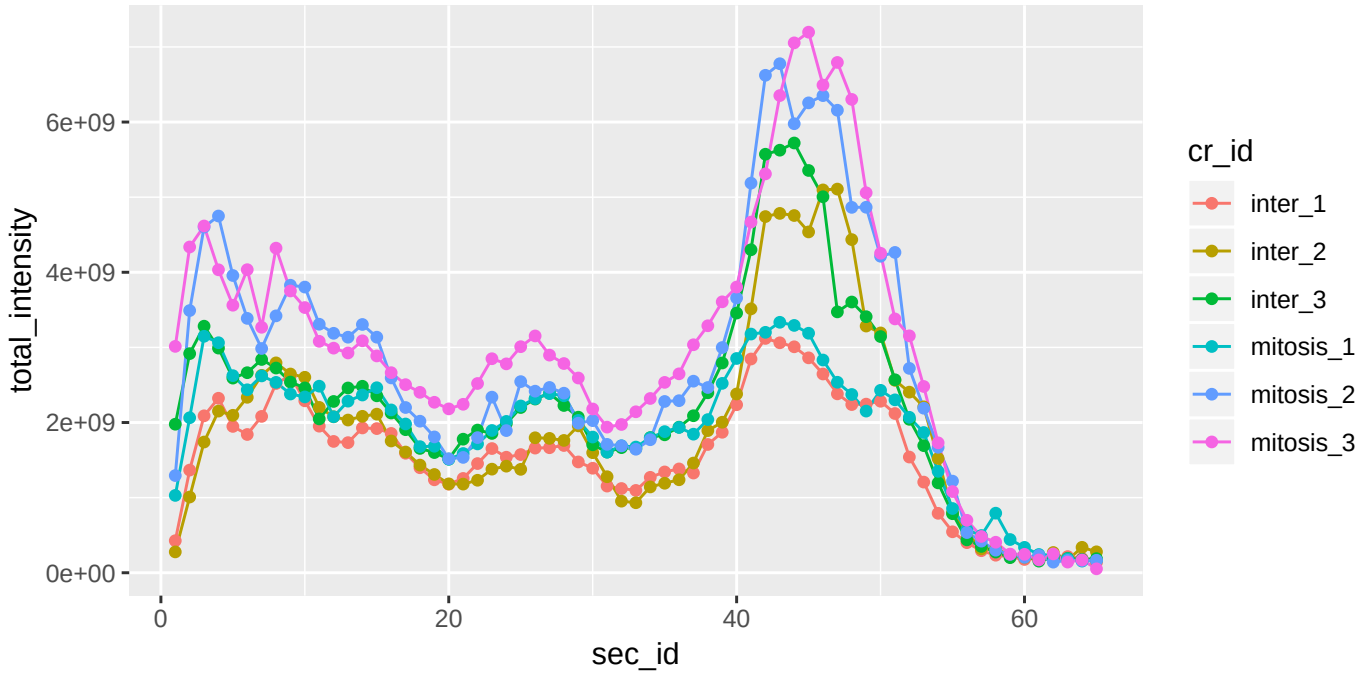
number of detections with the three replicates per condition. (**c**) The receiver-operating characteristic (ROC) illustrates the sensitivity vs. specificity of the different SECAT partial scores, the kickstart score and the integrated SECAT score. (**d**) The SECAT score histograms depict the dilution of the reference PPI network (true: CORUM) with false PPI (false: CORUM-inverted) for the benchmark (Methods). (**e**) The SECAT estimated q-value is accurate as evaluated against the ground truth for all levels of reference PPI network accuracy. For the dilution 1:0, no a priori false PPI are tested, thus the true FDR equals to 0. (**f**) The sensitivity of PPI detection in dependency of SECAT q-value for different levels of reference PPI network accuracy is depicted. Solid lines represent the true, whereas dashed lines represent the false PPI. (**g**) Consistency of PPI detection between replicates of the same conditions.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

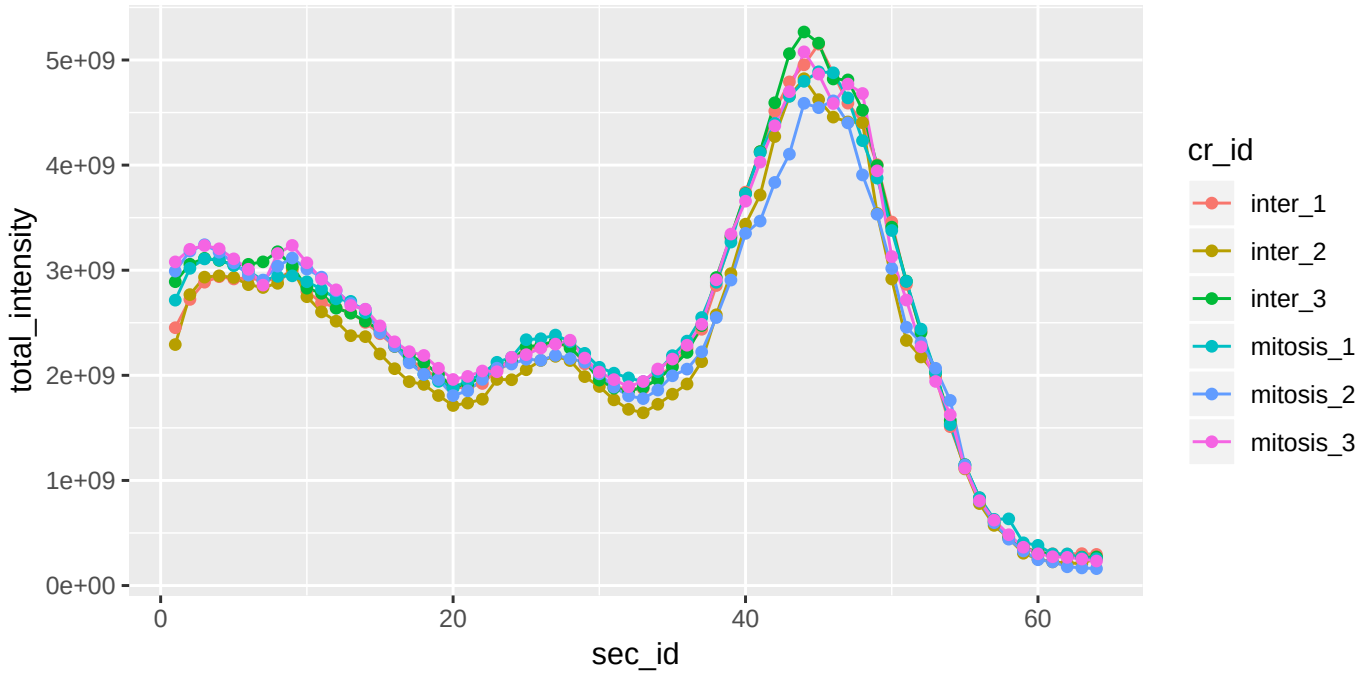


Supplementary Figure 2. Receiver operating characteristic for the comparison of SECAT with EPIC and CCprofiler using the HEK293-EG [3] dataset.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

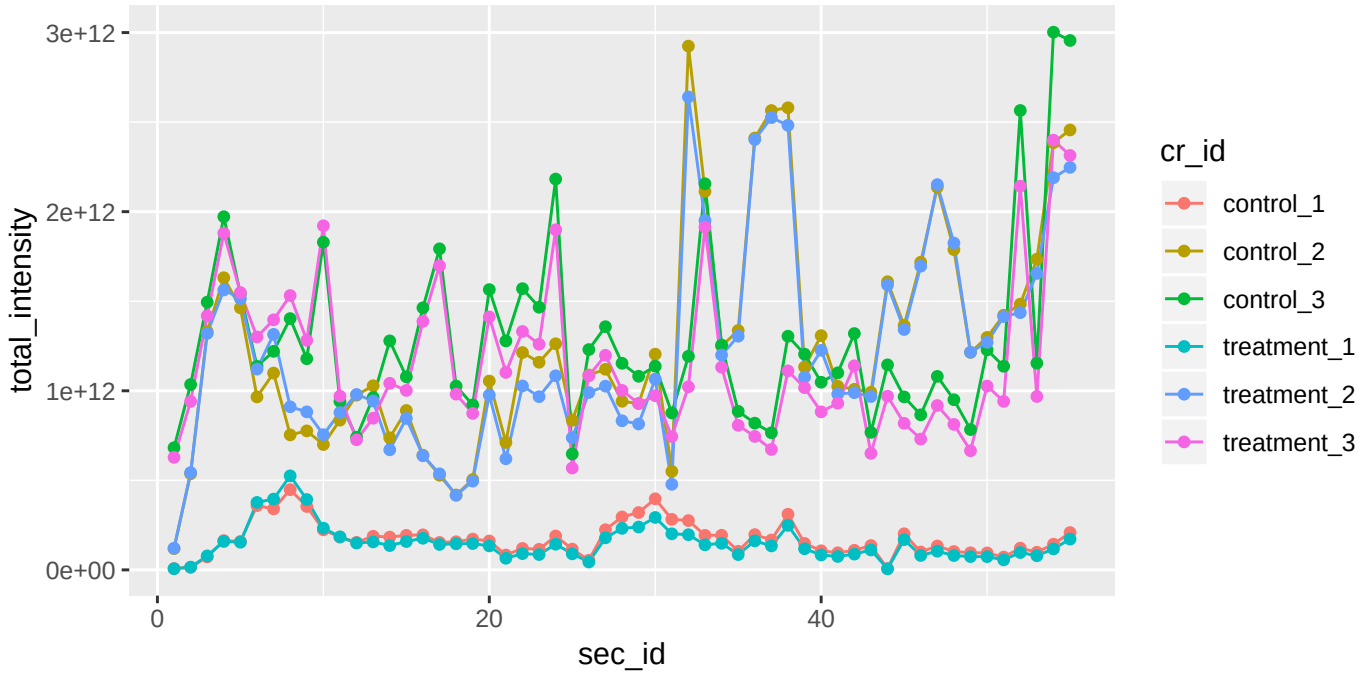


Supplementary Figure 3. Summarized raw peptide precursor intensities over the SEC profile for the HeLa-CC [38] dataset.

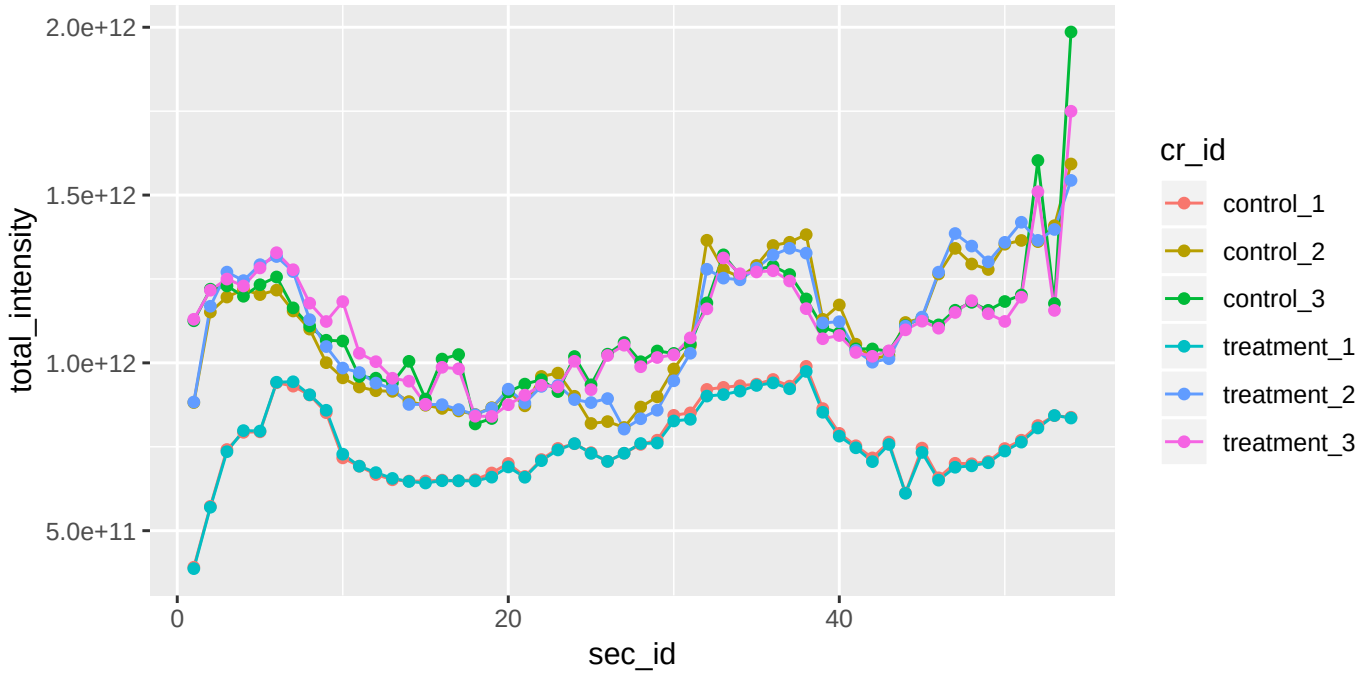


Supplementary Figure 4. Summarized normalized peptide precursor intensities over the SEC profile for the HeLa-CC [38] dataset.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 5. Summarized raw peptide precursor intensities over the SEC profile for the Jurkat-Fas [30] dataset.



Supplementary Figure 6. Summarized normalized peptide precursor intensities over the SEC profile for the Jurkat-Fas [30] dataset.

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
2. Campbell, I. D. The march of structural biology. *Nature Reviews Molecular Cell Biology* **3**, 377–381 (2002).
3. Heusel, M. *et al.* Complex-centric proteome profiling by SEC-SWATH-MS. *Molecular Systems Biology* **15**, e8438 (2019).
4. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
5. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
6. Gingras, A.-C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nature Reviews Molecular Cell Biology* **8**, 645–654 (2007).
7. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–643 (2006).
8. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell* **138**, 389–403 (2009).
9. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification–mass spectrometry data. *Nature Methods* **8**, 70–73 (2011).
10. Herzog, F. *et al.* Structural Probing of a Protein Phosphatase 2A Network by Chemical Cross-Linking and Mass Spectrometry. *Science* **337**, 1348–1352 (2012).
11. Bisson, N. *et al.* Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor. *Nature Biotechnology* **29**, 653–658 (2011).
12. Lambert, J.-P. *et al.* Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nature Methods* **10**, 1239–1245 (2013).
13. Collins, B. C. *et al.* Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. *Nature Methods* **10**, 1246–1253 (2013).
14. Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS). *Molecular & Cellular Proteomics* **14**, 120–135 (2015).
15. Varjosalo, M. *et al.* Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nature Methods* **10**, 307–314 (2013).
16. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).
17. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).

19. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).

20. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42**, D358–D363 (2014).

21. Drew, K. *et al.* Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology* **13**, 932 (2017).

22. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **494**, 127 (2013).

23. Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L. & Honig, B. PrePPI: A structure-informed database of protein-protein interactions. *Nucleic Acids Research* **41**, D828–33 (2013).

24. Garzón, J. I. *et al.* A computational interactome and functional annotation for the human proteome. *eLife* **5** (2016).

25. Foster, L. J. *et al.* A Mammalian Organelle Map by Protein Correlation Profiling. *Cell* **125**, 187–199 (2006).

26. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods* **9**, 907–909 (2012).

27. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).

28. 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. *Molecular & Cellular Proteomics* **12**, 3851–3873 (2013).

29. Wan, C. *et al.* Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015).

30. Scott, N. E. *et al.* Interactome disassembly during apoptosis occurs independent of caspase cleavage. *Molecular Systems Biology* **13**, 906 (2017).

31. Hu, L. Z. *et al.* EPIC: software toolkit for elution profile-based inference of protein complexes. *Nature Methods* **16**, 737–742 (2019).

32. Wan, C. *et al.* ComplexQuant: High-throughput computational pipeline for the global quantitative analysis of endogenous soluble protein complexes using high resolution protein HPLC and precision label-free LC/MS/MS. *Journal of Proteomics* **81**, 102–111 (2013).

33. 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. *Journal of Proteomics* **118**, 112–129 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

34. 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).

35. Picotti, P. & Aebersold, R. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nature Methods* **9**, 555–566 (2012).

36. Ting, Y. S. *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & Cellular Proteomics* **14**, 2301–2307 (2015).

37. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research* **47**, D559–D563 (2019).

38. Heusel, M. *et al.* A global screen for assembly state changes of the mitotic proteome by SEC-SWATH-MS. *bioRxiv* 633479 (2019).

39. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology* **3**, 1–2 (2011).

40. Hofmann, J. C., Husedzinovic, A. & Gruss, O. J. The function of spliceosome components in open mitosis. *Nucleus* **1**, 447–459 (2010).

41. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics* **48**, 838–847 (2016).

42. Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods* **8**, 430–435 (2011).

43. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* **32**, 219–223 (2014).

44. Albanese, D. *et al.* minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics* **29**, 407–408 (2013).

45. Teleman, J. *et al.* DIANA-algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* **31**, 555–562 (2015).

46. Rosenberger, G. *et al.* Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nature Methods* **14**, 921–927 (2017).

47. Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B. & Knijnenburg, T. A. Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics* **32**, i430–i436 (2016).

48. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **15**, 335–346 (2014).

49. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46**, D649–D655 (2018).

50. Gavet, O. & Pines, J. Progressive Activation of CyclinB1-Cdk1 Coordinates Entry to Mitosis. *Developmental Cell* **18**, 533–543 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/819755; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

51. Castro, A., Bernis, C., Vigneron, S., Labbé, J. C. & Lorca, T. The anaphase-promoting complex: A key factor in the regulation of cell cycle. *Oncogene* **24**, 314–325 (2005).

52. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology* **3**, 1–2 (2011).

53. Neumann, B. *et al.* Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721–727 (2010).

54. Rosenberger, G. *et al.* Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nature Biotechnology* **35**, 781–788 (2017).

55. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923–925 (2007).

56. Trabuco, L. G., Betts, M. J. & Russell, R. B. Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods* **58**, 343–348 (2012).

57. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794 (ACM, New York, NY, USA, 2016). URL `http://doi.acm.org/10.1145/2939672.2939785`.

58. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

59. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289–300 (1995).

60. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

61. Ballman, K. V., Grill, D. E., Oberg, A. L. & Therneau, T. M. Faster cyclic loess: Normalizing RNA arrays via linear models. *Bioinformatics* **20**, 2778–2786 (2004).

62. Kotlyar, M., Pastrello, C., Malik, Z. & Jurisica, I. IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Research* (2018).

63. Rosenberger, G., Ludwig, C., Röst, H. L., Aebersold, R. & Malmström, L. ALFQ: An R-package for estimating absolute protein quantities from label-free LC-MS/MS proteomics data. *Bioinformatics* **30**, 2511–2513 (2014).

64. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* 060012 (2016).

65. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–504 (2003).

66. Kucera, M., Isserlin, R., Arkhangorodsky, A. & Bader, G. D. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000Research* **5**, 1717 (2016).