**Analytical Guidelines for co-fractionation Mass Spectrometry Obtained through Global Profiling of Gold Standard *Saccharomyces cerevisiae* Protein Complexes**

**Authors**
Chi Nam Ignatius Pang, Sara Ballouz, Daniel Weissberger, Loïc M. Thibaut, Joshua J. Hamey, Jesse Gillis, Marc R. Wilkins, and Gene Hart-Smith

**In Brief**
To obtain guidelines for the best practice characterization of protein complexes using co-fractionation mass spectrometry (CF-MS), novel and published *Saccharomyces cerevisiae* CF-MS data sets are thoroughly evaluated using high proteome coverage libraries of yeast gold standard complexes. It is found that near-maximum recall of complexes can be achieved with fewer samples than published studies, several similarity scoring metrics – such as Spearman and Kendall correlations – are underutilized, and genomic data integration is generally unsuitable for predicting novel complexes in model organisms.

**Highlights**
- Guidelines for studying protein complexes via co-fractionation mass spectrometry.
- A novel procedure for profiling gold standard protein complexes in CF-MS data.
- Recommendations for efficient CF-MS fractionation collection.
- Scoring metric recommendations for precise and sensitive CF-MS data analysis.
Analytical Guidelines for co-fractionation Mass Spectrometry Obtained through Global Profiling of Gold Standard Saccharomyces cerevisiae Protein Complexes

Chi Nam Ignatius Pang¹, Sara Ballouz², Daniel Weissberger³, Loïc M. Thibaut⁴, Joshua J. Hamey¹, Jesse Gillis⁵, Marc R. Wilkins¹, and Gene Hart-Smith¹,6,*‡

Co-fractionation MS (CF-MS) is a technique with potential to characterize endogenous and unmanipulated protein complexes on an unprecedented scale. However this potential has been offset by a lack of guidelines for best-practice CF-MS data collection and analysis. To obtain such guidelines, this study thoroughly evaluates novel and published Saccharomyces cerevisiae CF-MS data sets using very high proteome coverage libraries of yeast gold standard complexes. A new method for identifying gold standard complexes in CF-MS data, Reference Complex Profiling, and the Extending ‘Guilt-by-Association’ by Degree (EGAD) R package are used for these evaluations, which are verified with concurrent analyses of published human data. By evaluating data collection designs, which involve fractionation of cell lysates, it is found that near-maximum recall of complexes can be achieved with fewer samples than published studies. Distributing sample collection across orthogonal fractionation methods, rather than a single high resolution data set, leads to particularly efficient recall. By evaluating 17 different similarity scoring metrics, which are central to CF-MS data analysis, it is found that two metrics rarely used in past CF-MS studies — Spearman and Kendall correlations — and the recently introduced Co-apex metric frequently maximize recall, whereas a popular metric — Euclidean distance — delivers poor recall. The common practice of integrating external genomic data into CF-MS data analysis is also evaluated, revealing that this practice may improve the precision and recall of known complexes but is generally unsuitable for predicting novel complexes in model organisms. If studying nonmodel organisms using orthologous genomic data, it is found that particular subsets of fractionation profiles (e.g., the lowest abundance quartile) should be excluded to minimize false discovery. These assessments are summarized in a series of universally applicable guidelines for precise, sensitive and efficient CF-MS studies of known complexes, and effective predictions of novel complexes for orthogonal experimental validation.

The physical interactions associated with proteins give rise to multiprotein complexes, other macromolecular assemblies and signaling pathways that are essential for cellular processes. These interactions can be graphed to produce complex networks, and potent insights into biological function can be gained by studying the topologies (1) and dynamics (2) of these networks. For this reason, one enduring goal of contemporary biology has been to comprehensively map the protein-protein interaction (PPI) networks that occur within organisms (1, 3). High-throughput methods that have been developed to meet this goal include yeast two-hybrid (Y2H), affinity purification-MS (AP-MS) and co-fractionation MS (CF-MS).

Of these high-throughput methods, CF-MS is unique in that it does not rely on heterologous expression or the genetic manipulation of cells or organisms. CF-MS has thus been able to predict endogenous and unmanipulated protein complexes on an unprecedented scale (4, 5) and to infer their PPIs (6). This has enabled notable insights into, for example, the evolution of eukaryotic protein complexes (5) and protein complex disassembly during apoptosis (7).

CF-MS involves extensive fractionation of cellular lysates — and the protein complexes therein — using one or more nonde-naturing separation techniques. The resulting fractionation profiles of individual protein complex subunits are measured using quantitative proteomics. As subunits of intact complexes will
co-fractionate, complexes can be bioinformatically predicted from these data using the correlations between fractionation profiles as a feature of central importance. Previous studies have typically made these predictions either within a machine learning framework (4, 5, 8–13), or by identifying interactions contained in existing protein interaction maps using techniques such as complex-centric profiling (14, 15).

Despite the successes of CF-MS, the method is yet to reach maturity. This is reflected in the diverse range of methodologies employed for CF-MS data collection and analysis (Table I). Regarding data collection, little is known about how different types and combinations of fractionation, and extents of fraction collection, impact upon the precision and recall of protein complex identification. This has resulted in CF-MS studies of dramatically different experimental cost. For example CF-MS data sets comprised of 40 size exclusion chromatography (SEC) fractions have previously been used for broad-scale identification of complexes in human U2OS cells (16), whereas in another study, performed on the lower complexity organism Desulfovibrio vulgaris, 5273 fractions collected using multiple levels of chromatography were used for a similar purpose (9). Regarding data analysis, a variety of correlation metrics, genomic data or other features have been employed in the machine learning classifiers (4, 5, 8–13) and complex-centric profiling workflows (14, 15) used to identify protein complexes. There is as yet no consensus on how effective these different features are. For example a recent re-analysis of a large-scale CF-MS data set indicated that, among the novel PPIs identified by the machine learning classifier employed in the original study, >85% were likely false positives and only <7% overlapped across independent screens (9). Together these examples speak to the urgent need for a better fundamental understanding of the recall and precision of protein complex identification using CF-MS.

Previous attempts to gain such understanding have relied on the use of gold standard protein complexes or PPIs to assess CF-MS data. Specifically, these gold standards have been used to provide an indication of co-purification accuracy during fractionation (9), profile CF-MS data sets (14–16), identify thresholds for predicting high confidence interactions from distance-based similarities between protein fractionation profiles (7, 17, 18), or for the training and testing of classifiers developed using supervised machine learning (4, 5, 8–13). The gold standards used for these purposes only represented limited portions of the human or bacterial complexomes under investigation, and often only exist under a restricted range of experimental conditions (19). Moreover they have often been employed under the assumption that they reflect the characteristics of protein complexes that are uniquely uncovered using CF-MS (4, 5, 7–13, 17, 18); an assumption that is difficult to test without very high proteome-coverage reference libraries of gold standards. As such, these gold standards have only provided limited insight into the best-practice collection and analysis of CF-MS data. Because extensive reference libraries of gold standard complexes and PPIs only exist for a few model organisms, very high proteome-coverage profiling of CF-MS data sets using gold standards has not yet been possible.

To overcome these limitations with the use of gold standards—and the resultant dearth in analytical guidelines for CF-MS—here we assess novel and published (5) CF-MS data sets of the model organism Saccharomyces cerevisiae, obtained using SEC and ion exchange chromatography (IEX) respectively. Saccharomyces cerevisiae has high proteome-coverage reference libraries of gold standard protein complexes and PPIs (20–22) derived from exhaustive AP-MS (22–24) and Y2H (25) surveys, which we exploit to thoroughly profile and evaluate these CF-MS data sets. By probing the impacts of the experimental design of the cell lysate fractionation and assessing different methods of fractionation profile similarity scoring, we uncover universally relevant guidelines for several critical aspects of CF-MS. These include guidelines for cost effective data collection and maximizing the precision and recall of protein complexes during data analysis. Differences between fractionation profiles, gene co-expression and Gene Ontology (GO) between gold standard and novel complexes are also probed, informing best-practice predictions of novel complexes. We complement these analyses by assessing published human SEC (16) and IEX (4, 5) CF-MS data sets using reference human complexes from the CORUM database (26), which further inform and reinforce our findings obtained from yeast. Together these analyses produce firm recommendations for the best-practice collection and analysis of CF-MS data, and a broad-scale understanding of the protein complexes that can be uniquely uncovered using CF-MS.

MATERIALS AND METHODS

Yeast Strain, Culture Conditions and Lysis—SEC CF-MS data were collected from Saccharomyces cerevisiae strain BY4741 (#YSC1052, Open Biosystems), grown at 30 °C in 300 ml YEPD media containing 2% (w/v) glucose, 2% (w/v) bactopeptone and 1% (w/v) yeast extract. Cells were harvested during mid-log phase growth (OD600 of 1.2). Harvested cells were washed with water and resuspended in 10 ml SEC mobile phase (50 mM NaCH3COO, 50 mM KCl, pH 7.2) complemented with cOmplete, EDTA-free Protease Inhibitor mixture (Roche) and PhosStop (Roche). To digest yeast cell walls, 600 lytic units of zymolase (Zymo Research) in 120 μL vendor-supplied resuspension buffer were added and the sample incubated at 30 °C until a stable OD600 of 3.5 was reached. The resulting spheroplasts were subjected to four 30 s rounds of sonication (amplitude 40%, 0.5 s pulse off/on) and centrifuged for 45 min at 14000 rpm at 4 °C. To enrich for protein complexes, lysates were then concentrated to 500 μL using 100 kDa MWCO filters (Sartorius Stedim) by centrifugation for 30 min at 3000 rpm.

Size Exclusion Chromatography—Lysates (2 replicate injections of 120 μl each) were loaded onto a 1290 Infinity UHPLC system (Agilent Technologies) and separated by SEC using a 600 × 7.8 mm Bio-Sep4000 column (Phenomenex). The mobile phase (described above) was run at a flow rate of 0.5 ml/min. For each injection 70 fractions were collected at a rate of 2/min from 20 to 55 min, and equivalent
### TABLE I

**Recent co-fractionation mass spectrometry (CF-MS) studies**

| Organism(s) Studied | Fractionation Method(s) Used | Number of Fractions Collected per Experiment | Similarity Scoring Metric(s) Used | Data Interpretation Methods | Ref. (Year) |
|---------------------|-----------------------------|---------------------------------------------|----------------------------------|----------------------------|-------------|
| Human (HeLa and HEK293 cells) | Primarily Ion exchange chromatography (IEX) | 43, 120 or 269/375† | Pearson correlation, normalized cross-correlation (NCC), Apex Location | Supervised Machine learning to infer PPIs (CORUM complexes for training and testing) and clustering to infer complexes | (4) (2012) |
| Human (HeLa cells) | size exclusion chromatography (SEC) | 50 | Euclidean distance | Fractionation profiles deconvolved into Gaussian curves, and features of curves compared to infer PPIs | (17) (2012) |
| Human Plasma | SEC, IEX, isoelectric focusing (IEF) | 17 (SEC), 23 (IEX), 20 (IEF) | Pearson correlation | Correlations from each fractionation method summed and used to infer PPIs | (58) (2013) |
| Human (U2OS cells) | SEC | 40 | Euclidean distance | Clustering to infer complexes; separate cross-comparisons with CORUM to identify high confidence complexes | (16) (2013) |
| Arabidopsis thaliana | SEC | 34 | Pearson correlation, NCC, Apex Location, 1 minus Euclidean distance | Clustering to infer complexes; supervised machine learning to infer PPIs (CORUM complexes for training and testing) and clustering to infer complexes | (59) (2014) |
| Caenorhabditis elegans, Drosophila melanogaster, Mus musculus, Strongylocentrotus purpuratus, human (CB660 and G166 cells), Xenopus laevis, Nematostella vectensis, Dictyostelium discoideum, Saccharomyces cerevisiae | IEX | 120 | Pearson correlation, NCC, Apex Location, 1 minus Euclidean distance | | |
| Human (U2OS cells) | SEC | 48 | Pearson correlation, NCC, Apex Location, Euclidean distance | Supervised machine learning to infer PPIs (CORUM complexes for training and testing) and clustering to infer complexes | (8) (2016) |
| Desulfovibrio vulgaris | Successive fractionation by IEX, HIC and SEC | 5,273 | Pearson correlation | Supervised machine learning to infer PPIs (curated *Escherichia coli* PPIs for training and testing) | (9) (2016) |
| Arabidopsis thaliana | SEC, sucrose velocity | 24/48 (SEC), 50 (sucrose velocity) | not specified | Fractionation profiles deconvolved into Gaussian curves; clustering of curves to infer putative complexes; Arabidopsis complexes predicted using orthologous CORUM complexes | (55) (2017) |
| human (Jurkat T cells) | SEC | 80 | Euclidean distance | Fractionation profiles deconvolved into Gaussian curves; features of curves compared to infer PPIs | (7) (2017) |
fractions from replicate injections were pooled to produce 70 samples of fractionated lysate.

**Tryptic Digestion**—To each sample of fractionated lysate a solution of 0.467 M DTT, 2.27 M chloroacetamide and 0.964 mM trypsin was added such that each sample had DTT and chloroacetamide concentrations of 47 mM and 236 mM respectively, and 3 mg trypsin. Samples were then incubated at 37 °C for 18 h. The resultant proteolytic peptide solutions were evaporated to dryness in a SpeedVac (Savant SPD1010, Thermo Scientific, Bremen, Germany), after which peptides were reconstituted in 1 ml 0.1% (v/v) heptafluorobutyric acid (pH 2.5) and C18 cleanups performed using Sep-pak cartridges (WAT054960) following manufacturer’s instructions. Eluted peptides from each cleanup were evaporated to dryness in a SpeedVac and reconstituted in 20 μL 0.1% (v/v) formic acid.

**Mass Spectrometry**—Proteolytic peptide samples were subjected to LC–MS/MS analysis on a Q Exactive Plus mass spectrometer.

### Table I—continued

| Organism(s) Studied | Fractionation Method(s) Used | Number of Fractions Collected per Experiment | Similarity Scoring Metric(s)* Used | Data Interpretation Methods | Ref. (Year) |
|---------------------|-----------------------------|---------------------------------------------|------------------------------------|----------------------------|-------------|
| *Trypanosoma brucei* | SEC, IEX                    | 48 (SEC), 96 (IEX)                          | Pearson correlation, NCC, Apex Location, Euclidean distance | Supervised machine learning to infer PPIs (CORUM and manually curated complexes for training and testing) and clustering to infer complexes | (11) (2017) |
| *Mus musculus*      | SEC                         | 80                                          | Pearson correlation, Euclidean distance, Peak Location, Co-apex | PrInCE workflow: fractionation profiles deconvolved into Gaussian curves; supervised machine learning to infer PPIs (CORUM for training and testing) | (60) (2018) |
| *Arabidopsis thaliana* | SEC, IEX                    | 38 (SEC), 65 (IEX)                          | Squared Euclidean distance         | Clustering to infer complexes, performed separately for IEX, SEC and IEX+SEC (reproducible profiles only) data sets | (61) (2019) |
| Human (HEK293 cells) | SEC                         | 81                                          | Pearson correlation                | CCprofiler workflow: detection of co-eluting subunits of known complexes using a sliding fraction window; FDR estimation using decoy complexes | (14) (2019) |
| Human (HeLa CCL2 cells) | SEC                         | 90                                          | Pearson correlation                | CCprofiler workflow using manually curated complexes for training and testing of the machine learning classifier; clustering to infer complexes | (15) (2019) |
| *Escherichia coli*  | SEC                         | 54                                          | Pearson correlation, Euclidean distance, Peak Location, Co-apex | PrInCE workflow using manually curated complexes for training and testing of the machine learning classifier; clustering to infer complexes | (12) (2019) |
| *Arabidopsis thaliana, Brassica oleracea, Glycine max, Cannabis sativa, Solanum lycopersicum, Chenopodium quinoa, Zea mays, Oryza sativa ssp. japonica, Triticum aestivum, Cocos nucifera, Ceratopteris richardii, Selaginella moellendorfii, Chlamydomonas reinhardtii* | SEC, IEX, IEF | 70 (SEC), 60 (IEX), 24 (IEF) | Pearson correlation, Spearman correlation, Euclidean distance, Bray-Curtis similarity, stationary cross-correlation | Assignment of proteins from each species into orthogroups; supervised machine learning to infer orthogroup PPIs (human CORUM and manually curated plant complexes for training and testing) and clustering to infer complexes | (13) (2020) |

*The term similarity scoring metric refers to the scoring (e.g. correlation) metric used to match fractionation profiles. Details for specific metrics are cited in Materials and Methods: Experimental design and statistical rationale.

† Multiple IEX columns were used, and different numbers of fractions were collected using each column.
(Thermo Scientific) interfaced with an UltiMate 3000 HPLC and auto-sampler system (Dionex, Amsterdam, The Netherlands). Peptides were separated by nano-LC and eluting peptides ionized using positive ion mode nano-ESI as described previously (27). Survey scans m/z 350–1750 (MS AGC = 1 × 105) were recorded in the Orbitrap (resolution = 70,000 at m/z 200). The instrument was set to operate in DDA mode, and up to the 12 most abundant ions with charge states of >+2 were sequentially isolated and fragmented via HCD using the following parameters: normalized collision energy 30, resolution = 17,500, maximum injection time = 125 ms, and MS2 AGC = 1 × 105. Dynamic exclusion was enabled (exclusion duration = 30 s).

MaxLFQ of Novel and Published co-Fractionation Mass Spectrometry Raw Files—To generate SEC fractionation profiles for individual yeast protein complex subunits, LC-MS/MS raw files were analyzed using MaxQuant (version 1.6.2.10) (28). Sequence database searches were performed using Andromeda (29) and the MaxLFQ algorithm (30) was used to quantify proteins across fractions. The following parameters were employed: precursor ion and peptide fragment mass tolerances were ±4.5 ppm and ±20 ppm respectively; carbamidomethyl (C) was included as a fixed modification; oxidation (M) and N-terminal protein acetylation were included as variable modifications; enzyme specificity was trypsin with up to two missed cleavages; only S. cerevisiae sequences in the Swiss-Prot database (February 2017 release, 553,655 sequence entries) were searched; the minimum peptide length was set as 7; the “match between runs” feature was enabled; and MaxLFQ analyses were performed using default parameters with “fast LFO” enabled. Protein and peptide false discovery rate thresholds were set at 1%. Only fractionation profiles obtained from proteins identified from ≥2 peptides and with nonzero MaxLFQ values in ≥2 fractions were subjected to downstream analysis. All MaxLFQ values were added with a pseudo-count of one and log-transformed (base 10).

To generate fractionation profiles from publicly available CF-MS data which are amenable to MaxLFQ analysis following the above procedures, additional MaxQuant analyses were performed on raw files associated with the following 3 CF-MS experiments: S. cerevisiae lysate separated across 108 IEX fractions, as performed by Wan et al. (5) (PRIDE data identifier PXD002327); human G166 glioma stem cell lysate separated across 108 IEX fractions, as performed by Havugimana et al. (4) (PRIDE data identifier PXD002322); and human U2OS lysate separated across 40 SEC fractions, as performed by Kirkwood et al. (biological replicate 3) (16) (PRIDE data identifier PXD001220). For each set of raw files, MaxQuant analyses were performed as above with the following alterations: precursor ion and peptide fragment mass tolerances were ±4.5 ppm and ±0.5 Da respectively, and only human sequences in the Swiss-Prot database were searched when analyzing raw files associated with human cell lines.

To test the effects of collecting fewer fractions during CF-MS data collection, simulated fractionation profiles with reduced resolution were generated from the above results. For individual proteins in each CF-MS data set, this involved redistributing the MaxLFQ values from the original number of fractions into a smaller number of fractions using the proportional scaling method detailed in supplemental Fig. S1.

To identify and characterize Gaussian peaks in fractionation profiles the PrInCE software (10) was used, with up to five peaks identified for each protein.

**Experimental Design and Statistical Rationale**

**Gold Standard Protein Complex Data sets—**To interpret fractionation profiles produced in S. cerevisiae CF-MS experiments, the reference library of gold standard S. cerevisiae protein complexes published by Benschop et al. (21) was used. This library contains 1,881 proteins from 518 complexes, and is referred to herein as the Benschop reference set. To interpret fractionation profiles produced in human cell-line CF-MS experiments, the reference library of gold standard human protein complexes in CORUM version 02.07.2017 (31) was used. This library contains 3241 proteins from 2389 complexes.

**Identifying Gold Standard Protein Complexes through Reference Complex Profiling—**To find evidence for gold standard protein complexes in CF-MS data sets, significantly related fractionation profiles were identified using the randomization procedure illustrated in Fig. 1, hereafter referred to as Reference Complex Profiling. Reference Complex Profiling first calculates mean fractionation profile similarity scores for reference library complexes in which two or more subunits are identified in the CF-MS data set. The fractionation profile similarity scores for all unique pairs of proteins in each query complex are calculated using a similarity scoring metric (e.g., Pearson correlation), and the mean of these scores determined. Secondly, a bootstrapping method tests whether the observed mean fractionation profile similarity score for each query protein complex is statistically significant as compared with the background distribution, generated from 1000 random protein complexes. Each random complex is generated by randomly sampling the same number of proteins observed in the query complex from proteins identified in the CF-MS data set. The bias corrected p-value of each query complex is calculated as a ratio, with a numerator of one plus the number of times the observed mean fractionation profile similarity score of the query complex is larger than the mean fractionation profile similar scores of 1,000 random protein complexes, and a denominator of 1001 (32). Minor alterations to this calculation are applied when using the following similarity scoring metrics: Euclidean distance, maximum cumulative distance between two scaled vectors (DMax), Co-apex and Peak Location (each detailed below). For these metrics, the observed score for a co-fractionating group of protein complex subunits should generally be lower, rather than higher, than the score of a random complex.

When analyzing stand-alone CF-MS data sets, p-values obtained from this randomization procedure were adjusted using the Benjamini-Hochberg procedure to control the false discovery rate (33), and reference complexes under query with bootstrapped p-values less than 0.05 were deemed significant. When co-analyzing SEC and IEX CF-MS data sets, Fisher’s method (34) was used to combine p-values. The natural log of SEC and IEX p-values, prior to false discovery rate adjustment, were summed and multiplied by negative two, giving a Chi-square statistic with two degrees of freedom as follows:

\[
\chi^2_2 = -2 \log(P_{SEC} + \log(P_{IEX}))
\]

Chi-square statistics were converted to p-scores using the Chi-square distribution function in R. When a reference complex was only queried in one of the SEC or IEX data sets, the p-score was the p-value from the data set in which the complex was queried. Resulting lists of p-scores were adjusted using the Benjamini-Hochberg procedure (33), and reference complexes with bootstrapped p-scores less than 0.05 were deemed significant.

**Similarity Scoring Metrics Used for Reference Complex Profiling—** Seventeen different similarity scoring metrics were used to profile CF-MS data sets using the above procedures. Scoring metrics were calculated using functions from publicly available R libraries, or custom R scripts available from: https://github.com/IgnatiusPang/ReferenceComplexProfiling. The metrics covered four broad categories: correlation-based (Pearson correlation, Spearman correlation, Kendall correlation (35) and normalized cross-correlation (NCC) (8)); distance-based (Euclidean distance, Distance correlation (36), Jaccard...
distance and DMax); mutual information-based (Mutual Information (MI), Biased Corrected Mutual Information (BCMI) (37), Maximum Information Coefficient (MIC) (38), Total Information Coefficient (TIC) (39), Generalized Mean Information Coefficient (GMIC) (40) and Randomized Information Coefficient (RIC) (41)); and peak-based (Apex Location (4), Peak Location (10) and Co-apex (10)). Details for how each metric was applied are provided in the Supporting Information.

Gene-Gene Networks and Reference Gene Annotations—To test the effects of integrating CF-MS data sets with external genomic data, CF-MS networks were evaluated against reference gene annotation sets (e.g. GO). Weighted CF-MS networks were generated using the Spearman correlations between fractionation profiles in individual CF-MS data sets. The nodes of these networks were the individual genes/proteins associated with each fractionation profile. Edge scores from each network were converted into ranks, with ties averaged. The resulting ranks were divided by the maximum rank in the network to scale values between 0 and 1. Reference gene annotation sets were obtained from the GO Consortium (version September 2017) (42, 43) for yeast (6,392 genes across 8,672 GO terms) and human (19,030 genes across 21,946 GO terms). Additional reference gene annotation sets, comprised of high confidence protein-protein interactions, were obtained from Pang et al. for yeast (44) and Huttlin et al. for human (45).

Measuring Network Information Content Using ‘Guilt-by-Association’—CF-MS networks were evaluated against reference gene annotation sets using the Extending ‘Guilt-by-Association’ by Degree (EGAD) R package (46). EGAD reports AUROCs as a performance metric, which ranges between 0 and 1, with 0.5 random and 1 perfect. EGAD evaluations were performed on both unaltered and merged networks. For merged networks, gene co-expression networks for yeast (47) or human (48), or high confidence PPI networks for yeast (44) or human (45), were merged with CF-MS networks. Because PPI networks consist of binary edge scores, indirect connections were modeled by adding edges between node pairs with minimum path length less than seven, weighted by the inverse of the minimum path length. Networks were merged by taking the union of networks and summing the edge score of overlapping edges, before applying the rank conversion and scaling described above. Only genes or gene products common in all yeast or all human networks (CF-MS, gene co-expression and PPI) were used in merged networks, and all network comparisons were performed on networks containing the same subset of genes or proteins. Networks comprised of protein interactions were not evaluated against reference gene annotation sets comprised of protein interactions.

Additional sub-networks, comprised of well-correlated fractionation profiles (Spearman correlation > 0.8) in CF-MS data sets sorted into networks of either known or putative novel interactions, were analyzed alone and merged with co-expression networks. Interactions with either direct or indirect evidence in the BioGrid (49, 50) or Interactome3D (51) databases were considered known, with all other interactions considered putative novel. For example, if interactions between both proteins A and B and proteins B and C are known based on direct evidence, interactions between proteins A and C are known based on indirect evidence. Sub-network comparisons were performed on networks containing the same subset of genes or proteins.

RESULTS

The below results begin with a general analysis of the breadth and reliability of the two yeast and two human CF-
MS data sets under investigation. Following this, results describing systematic assessments of these data sets using reference protein complexes and PPIs are presented. Experiments performed using Reference Complex Profiling are firstly shown. These experiments provide insight into two of the most fundamental questions regarding CF-MS data analysis and collection: how should the similarity of fractionation profiles be measured (e.g., correlated) during data analysis? and, how should the fractionation of cell lysates be designed during data collection? Assessments performed using other methods, for example EGAD (46), are then presented. These latter experiments provide insight into the usefulness of features other than fractionation profile similarity scoring for CF-MS data analysis, particularly in the context of identifying known versus novel complexes.

Subunits of Gold Standard Protein Complexes co-Fractionate in CF-MS Data Sets—The two yeast CF-MS data sets under investigation were obtained from lysates of S. cerevisiae strain BY4741 fractionated using SEC, and S. cerevisiae strain W303 fractionated using mixed-bed IEX (5). Two additional CF-MS data sets, obtained from human cell lysates, were also investigated: one from U2OS cells fractionated using SEC (16), and another from human G166 glioma stem cells fractionated using mixed-bed IEX (4). These are hereafter respectively referred to as the yeast SEC, yeast IEX, human SEC and human IEX CF-MS data sets.

Fig. 2 shows Kernel Density Plots of the Pearson correlations between fractionation profiles observed in these data sets. For each data set, these are separated into two distributions: one showing the correlations between shared subunits of reference complexes, and another showing the correlations between all other pairs of fractionation profiles. The fractionation profiles of shared subunits of reference complexes correlate to an overall greater extent than other pairs of fractionation profiles. This suggests that for each CF-MS data set, reference complexes were present in the samples under analysis and successfully fractionated.

The fractionation profiles within each CF-MS data set were generated using equivalent sequence database searching and label-free quantification methods, and only fractionation profiles with nonzero MaxLFQ values in ≥2 fractions were kept (as detailed in Materials and Methods). This resulted in the following numbers of fractionation profiles per CF-MS data set: 1280 for the yeast SEC data set, 1894 for the yeast IEX data set, 3,935 for the human SEC data set, and 2922 for the human IEX data set.

Together these results indicate that the data sets under analysis are suitable for assessing different methods of broad-scale protein complex identification using CF-MS, and that the profiling of reference protein complexes and PPIs can play a role in these assessments.

Performance Characteristics of Different Fractionation Profile Similarity Scoring Metrics—The results described in Fig. 3 provide insight into the most fundamental aspect of CF-MS data analysis: measuring the similarity of fractionation profiles. Fig. 3A shows the numbers of significant gold standard complexes identified from each CF-MS data set from two or more identified subunits using Reference Complex Profiling. Each data set was systematically profiled 17 times using Reference Complex Profiling; once for each similarity scoring metric listed in Materials and Methods. This enabled multiple methods for performing fractionation profile similarity scoring to be directly compared. Because these comparisons are not reliant on depth of complexome coverage, results from the profiling of yeast and human CF-MS data sets using the Benschop and CORUM reference sets of gold standards, which cover relatively high and low portions of the yeast and human proteomes respectively, are presented alongside one another.

Fig. 3A reveals that some scoring metrics are more widely suitable for CF-MS data analysis than others. It is notable that relative scoring metric performances remain broadly consistent across the four diverse CF-MS data sets studied here. Three of the four correlation-based metrics under analysis, Pearson, Spearman and Kendall, and the peak-based Co-apex metric are consistently among the top performing metrics when performance is judged on the number of statistically significant gold standard complexes identified. Three of the four distance-based metrics – Distance correlation, Jaccard distance and DMax – and the peak-based Apex Location metric also generally perform well, whereas mutual information-based metrics usually identify slightly fewer complexes on average in comparison. In contrast to the aforementioned metrics, NCC, Euclidean distance and Peak Location consistently identify a low number of complexes. Strikingly, a commonly used scoring metrics in CF-MS data analysis – Euclidean distance (see Table I) – returns no significant hits from the present data sets. Together these results indicate that some scoring metrics are more sensitive than others when applied to the analysis of entire CF-MS data sets.

To provide additional insight into these differences in scoring metric sensitivity, mean score distributions for complexes subjected to Reference Complex Profiling were compared across scoring metrics. Supplemental Fig. S2 shows these distributions for both significant and nonsignificant complexes. These distributions reveal that some scoring metrics, such as Peak Location and Euclidean distance, identify few significant protein complexes because they generally struggle to match fractionation profiles belonging to protein complex subunits. In contrast, other scoring metrics such as NCC, and MIC when applied to the yeast data sets, identify few significant protein complexes because they match fractionation profiles relatively indiscriminately.

To gain further insight into why these differences in scoring metric performance are observed, the fractionation profiles associated with significant protein complexes were analyzed. Specifically the peak heights, peak widths and numbers of
peaks per fractionation profile identified using PrInCE (10), together with fractionation profile lengths, were compared for the subunits of significant complexes returned from each scoring metric and category of scoring metric. In addition, the numbers of subunits per significant complex were also compared. Fig. 3B shows instances in which these
comparisons reveal significant differences. The box plots on the left show that mutual information-based metrics have a greater preference for matching CF-MS fractionation profiles with multiple peaks than nonmutual information-based metrics. These differences are significant for multiple individual mutual information-based metrics, as shown in supplemental Fig. S3. The box plots on the right show that, relative to other scoring metrics, the Peak Location metric has a greater preference for matching fractionation profiles with narrow peaks. Supplemental Fig. S4 shows these comparisons for each individual scoring metric under investigation. Aside from these observations, none of the other features of CF-MS fractionation profiles analyzed here, nor the numbers of subunits per significant complex, consistently differentiate scoring metrics or categories of scoring metrics across CF-MS data sets (see supplemental Figs. S5–S8). Together these results reveal that certain scoring metrics can preferentially match CF-MS fractionation profiles with certain characteristics, and that scoring metrics with low overall sensitivities may prefer to match fractionation profiles with atypical characteristics.

In addition to the total numbers of complexes identified using each scoring metric, complexes uniquely identified using each scoring metric were also investigated (see supplemental Figs. S9–S12). This reveals that even though some scoring metrics are less sensitive overall than others, most scoring metrics or categories of scoring metrics are capable of identifying complexes that others cannot. Fig. 3C provides two examples. The example on the left, taken from the yeast IEX CF-MS data set, shows fractionation profiles for subunits of the yeast Asn1-Asn2 isozyme complex. This complex was only identified using mutual information-based metrics (MI, BCMI and RIC). These fractionation profiles contain multiple peaks according to PrInCE, several of which are observed in...
later IEX fractions at low peak heights and are thus illustrative of the trends observed for mutual information-based metrics presented in Fig. 3B. The example on the right is taken from the same data set and shows fractionation profiles for subunits of the yeast Ski complex, which was uniquely identified using the Peak Location metric. These fractionation profiles are illustrative of the trends observed for the Peak Location metric presented in Fig. 3B. That is, the co-eluting peaks associated with this complex are particularly narrow. These examples illustrate that scoring metrics with relatively low overall sensitivities can have unique advantages for identifying particular types of complexes, such as complexes that elute across narrow time windows during chromatographic separation.

Taken together, these results suggest that Pearson, Spearman and Kendall correlations and the Co-apex metric have particularly broad-scale utility for CF-MS data analysis. However, they also reveal that other scoring metrics can provide complementary results. The implications of these findings are elaborated on in the Discussion.

Performance Characteristics of Stand-Alone and Orthogonal Fractionation—Having studied the most fundamental aspect of CF-MS data analysis in the previous section, this section turns attention to the most fundamental aspect of CF-MS data collection: the design of the fractionation. A chief consideration in this is to minimize the co-fractionation of unrelated proteins. This is influenced by the number of fractions collected, sample complexity, and the resolving power of the employed fractionation method. Another consideration is experimental cost. Because each collected fraction requires LC–MS/MS analysis, the fewer the number of fractions collected, the lower the cost. Experimental design of CF-MS data collection involves balancing these two considerations.

Fig. 3A indicates that for the present fractionation methods, which are identical in the yeast and human IEX data sets and broadly similar in the yeast and human SEC data sets, the increase in complexity from yeast to human samples does not dramatically increase the extent to which unrelated proteins produce matching fractionation profiles. This is apparent when considering the numbers of significant complexes identified. There is an approximately 10-fold increase in the number of significant complexes identified in human compared with yeast when equivalent scoring metrics and fractionation techniques are compared. This is even though the library of gold standard human complexes used to profile these data sets is only 5-fold larger than the yeast library (2,389 versus 518 complexes respectively). This suggests that complexes remain well resolved in the human CF-MS data sets relative to the equivalent yeast data sets. The increased complexity of the human samples relative to yeast samples did not necessitate an increase in the numbers of fractions collected.

Fig. 4A extends these observations. It shows the numbers of significant complexes identified following Reference Complex Profiling of the present CF-MS data sets, after the effects of reduced fraction collection had been simulated following the procedures described in Materials and Methods. (Results for the human SEC data set are not shown as this data set was collected from only 40 fractions.) In the yeast IEX data set, on average fewer significant complexes are identified per scoring metric when fraction collection is reduced from 108 to 70 samples, and then again to 40 samples. However, these decreases are not statistically significant. In the yeast SEC data set, overall, the simulated decrease in fraction collection results in no pronounced effects on the numbers of significant complexes identified. Strikingly this is also the case for the human IEX data set, which was collected from a relatively complex sample and in which the simulated reduction in fraction collection was dramatic. Together these results indicate that even when the co-elution of unrelated proteins is pronounced, protein complexes generally remain well resolved (i.e. produce distinct fractionation profiles) using the present fractionation methods over the ranges of fraction collection investigated here. They reinforce the above suggestion that increasing the numbers of fractions collected from the present samples is unlikely to result in the identification of substantially more complexes.

The above results consider the use of stand-alone fractionation methods for CF-MS data collection. Results pertaining to the combined use of orthogonal fractionation methods are presented in Figs. 4B and 4C. Fig. 4B shows the numbers of significant complexes identified from stand-alone and combined analyses of the yeast SEC and IEX CF-MS data sets across different simulated amounts of fraction collection. Two methods of combined analysis were undertaken. The first used Fisher’s Exact test to combine p-values following the procedures in Materials and Methods, and the second pooled the significant complexes identified from stand-alone analyses of the SEC and IEX CF-MS data sets. Both methods assume that although the samples used to create the SEC and IEX data sets were prepared from different yeast strains and subject to nonidentical lysis conditions, they nonetheless have a substantial number of gold standard protein complexes in common. Fig. 4B reveals that 30–50% percent more significant protein complexes are identified in combined analyses of SEC and IEX CF-MS data sets than in stand-alone analyses of these data sets, when analyses associated with equivalent numbers of fractions are compared. Overall the co-analysis of SEC and IEX CF-MS data sets using Fisher’s Exact test identifies the most significant complexes per fraction collected. Fig. 4C provides an example of a complex uniquely identified using Fisher’s Exact test: the yeast CURI complex. In stand-alone analyses of SEC and IEX CF-MS data sets using Spearman correlations, the fractionation profiles of the subunits of the CURI complex cannot be significantly differentiated from those of randomly generated complexes. However, Fisher’s Exact test reveals that the
combined orthogonal evidence for this complex in the SEC and IEX CF-MS data sets is statistically significant.

Together the above results indicate that for a set threshold of precision, recall of gold standard protein complexes is similar across broad ranges of fraction collection for most scoring metrics. This is true across the range of samples and fractionation methods studied here. They also indicate that CF-MS experiments are most cost effective if increases in fraction collection are distributed across orthogonal separation methods, rather than increased for any one stand-alone method. The implications and limitations of these findings are elaborated on in the Discussion.

**Fractionation Profile Characteristics for Subunits of Known versus Putative Novel Complexes**—The above analyses were limited to the high confidence profiling of gold standard protein complexes, in which stand-alone scoring metrics alone were
used for characterization. Additional complexes are likely to be present in the CF-MS data sets under investigation. The results in this section provide insight into how CF-MS fractionation profiles alone may assist in identifying these additional complexes, both known and novel.

Evidence indicating that the present CF-MS data sets likely include both known and novel complexes is contained in Fig. 5A. This figure identifies all fractionation profiles capable of being well correlated (Spearman correlations $> 0.8$) with other fractionation profiles; that is, fractionation profiles with likelihoods of being associated with protein complexes. For both yeast CF-MS data sets, Fig. 5A shows bar charts grouping these fractionation profiles into the following categories: those associated with known interactions only, those associated with both known and putative novel interactions, and those associated with putative novel interactions only. The proportions of fractionation profiles which cannot be well correlated are also shown. Comparative results from the 2 human CF-MS data sets are shown alongside those of yeast. It can be seen that, relative to human, high proportions of the fractionation profiles in the yeast CF-MS data sets are only associated with known interactions. This is expected because the reference libraries of PPIs in yeast are high proteome coverage relative to human. However, despite

**S. cerevisiae** being a particularly highly benchmarked model organism, many of the yeast CF-MS fractionation profiles are associated with putative novel interactions and ~40% cannot be interpreted using existing reference libraries of PPIs.

Fig. 5B indicates that fractionation profiles associated with putative novel interactions have distinct characteristics to those derived from known interactions. In yeast, it can be seen that the putative novel complexes associated with these interactions are significantly smaller (i.e. elute in later SEC fractions) and elute significantly earlier during mixed-bed IEX than known complexes when comparing the centers of peaks observed in fractionation profiles. Proteins only associated with putative novel interactions also appear to be of low abundance relative to those only associated with known interactions. That is, their relative peak heights and widths are lower on average, suggesting that they elute at low abundance across relatively few fractions. These observations are statistically significant in the yeast SEC and yeast IEX data sets, respectively. It is also observed that the subunits of putative novel complexes produce fractionation profiles with significantly fewer peaks than those of known complexes; an observation which is consistent with these subunits being of low abundance. All these observations are strongly
reinforced by the human data sets, which show identical trends to those in yeast.

Together these findings indicate that CF-MS data sets of model organisms contain bodies of evidence for both known and novel PPIs, and that the CF-MS fractionation profiles of these two classes of PPI have distinct characteristics. They identify possible avenues toward improved characterization of known and novel protein complexes from the distinct features of their CF-MS fractionation profiles alone, as elaborated on in the Discussion.

Assessment of Genomic Data Integration in CF-MS Data Analysis—The results in the above sections relate to the identification of protein complexes solely using CF-MS fractionation profiles. However, most CF-MS studies seek to predict protein complexes using features beyond those related to fractionation profiles; a process which has been termed genomic data integration (52). The results in this section assess the utility of genomic data integration via EGAD evaluations (46) of the present CF-MS data sets. They assess two types of genomic data—gene co-expression and high confidence PPI—which are used in many of the supervised machine learning methods of Table I, and the GO, which is frequently used to define cluster cutoff distances in the clustering methods of Table I.

The EGAD results for the yeast CF-MS data sets are shown in Fig. 6, with comparative results from the human CF-MS data sets shown alongside. These experiments provide an indication of the coherence of the information contained in the networks under investigation. That is, they show the relative capacities of network neighbors (e.g. well correlated fractionation profiles in CF-MS data sets) to predict shared GO, with AUROC scores > 0.5 indicating better than random performance. Similar experiments showing the relative capacities of network neighbors to predict high confidence PPIs are shown in supplemental Fig. S13.

The experiments of Fig. 6A reveal that networks of well correlated fractionation profiles in yeast CF-MS data sets can predict shared GO (blue bars). However, these predictive capacities are lower than those of yeast gene co-expression or high confidence PPI networks (purple bars). If the yeast SEC and IEX CF-MS data sets are merged following the procedures described in Materials and Methods, the information contained in the resultant network is no more coherent than in the stand-alone networks (pink bar). Similarly, if correlated fractionation profiles in yeast CF-MS data sets are merged with yeast gene co-expression (47) or high confidence PPI networks (44), the resultant AUROC scores reflect those of the stand-alone gene co-expression or high confidence PPI networks (light brown bars). This contrasts with the results observed when yeast gene co-expression and high confidence PPI networks are merged. The network neighbors in this merged network have a higher capacity to predict shared GO (dark brown bar) than in the stand-alone networks. All the above trends are also observed when using high confidence yeast PPIs as the reference set instead of GO (supplemental Fig. S13).

Inspection of Figs. 6A and S4 shows that the above results from yeast are broadly similar to those observed, with some minor differences. Specifically, neighbors in human gene co-expression (48) or high confidence PPI (45) networks only predict shared GO to a marginally higher extent than those in human CF-MS networks (purple and blue bars in Fig. 6A respectively), whereas the human gene co-expression network predicts high confidence human PPIs (purple bars in supplemental Fig. S13) to a lesser extent than the human CF-MS networks (blue bars in supplemental Fig. S13). Another difference is that when gene co-expression or high confidence PPI networks are merged with CF-MS networks in human, marginal improvements in AUROC scores are sometimes observed relative to the stand-alone networks (light brown bars). Altogether, however, these differences between yeast and human are proportionally minor.

Taken together, the results of Figs. 6A and supplemental Fig. S13 indicate that both GO and high confidence PPIs can assist in the interpretation of CF-MS data sets. However, they also suggest that, for the purposes of predicting protein complexes, there is little additive value in incorporating gene co-expression or high confidence PPI information in CF-MS data analysis.

Fig. 6B provides further insight into these findings by studying known and putative novel PPIs separately. In the yeast CF-MS data sets, only the well correlated fractionation profiles associated with known PPIs can predict shared GO (light blue bars); those associated with putative novel PPIs have almost no predictive capacity (dark blue bars). In addition, merged networks associated with both known and putative novel PPIs produce AUROC scores that reflect those of the stand-alone yeast gene co-expression network (light brown bars and purple bar respectively), which is consistent with Fig. 6A. The results observed in yeast are similar to but more pronounced than those observed in human.

Taken together, these experiments demonstrate that GO can play a role in identifying reference protein complexes in CF-MS data sets, whereas gene co-expression and high confidence PPI data can be used to increase the coherence of information generated from CF-MS data analysis. However, these experiments also demonstrate that GO, gene co-expression data and high confidence PPI data can only play a limited role in the prediction of protein complexes and PPIs, as elaborated on in the Discussion.

DISCUSSION

The above results provide new insights into how different CF-MS approaches impact the precision and recall of protein complex identification. The below sections firstly place these insights into the context of the existing CF-MS literature. Following this, a series of recommendations for best practice
A. EGAD evaluations of CF-MS networks

![Graph showing GO evaluations for Yeast and Human](image)

- CF-MS network
- Merged networks (CF-MS)
- Merged networks (other)
- Other network

B. EGAD evaluations of CF-MS networks (known versus putative novel interactions)

![Graph showing GO evaluations for Yeast and Human](image)

- CF-MS network (known interactions)
- CF-MS network (putative novel interactions)
- Other network

- Merged networks (CF-MS known interactions + other)
- Merged networks (CF-MS putative novel interactions + other)
Different Fractionation Profile Similarity Scoring Metrics Have Different Performance Characteristics—Past CF-MS studies have employed a variety of scoring metrics to assess the similarity of CF-MS fractionation profiles. Pearson correlation and Euclidean distance are particularly commonly used, and it is also common practice for machine learning classifiers to employ a panel of scoring metrics as features (Table I). There is, however, little consensus on how well these scoring metrics perform relative to one another. Salas et al. recently provided some insight into this question by comparing the precision of interactions identified using 11 individual scoring metrics, as estimated by PrInCE (53). This identified that some scoring metrics are often more precise than others, but that results differ across CF-MS data sets.

The results described in Fig. 3 substantially extend our knowledge of best-practice fractionation profile similarity scoring. They provide a measure of relative scoring metric sensitivity, while assessing performance from the perspective of characterizing entire protein complexes rather than individual interactions only. Moreover, in addition to the scoring metrics listed in Table I, they assess a range of scoring metrics that have not previously been applied to CF-MS data analysis.

In considering the implications of these results, it is useful to emphasize that protein complexes obtained from cellular lysates have diverse biophysicochemical properties and may be involved in diverse nonprotein interactions (e.g. RNA interactions). In any CF-MS experiment, different protein complexes will therefore be resolved to different extents during fractionation, producing data sets with ranges of fractionation profile characteristics. This corroborates the present finding that even though some scoring metrics have broader scale utility than others, no single scoring metric is optimal for every protein complex contained in a CF-MS data set.

The results of Fig. 3 support the current widespread use of Pearson correlation as a broadly suitable scoring metric for CF-MS data analysis. In addition, Spearman and Kendall correlations—which are not commonly used in CF-MS data analysis—and the recently introduced Co-apex metric (10) are found to have similar broad-scale utility. The results do not, however, support the practice of using Euclidean distance as the sole scoring metric in CF-MS studies designed to maximize recall of protein complexes.

One of the findings relating to Fig. 3—namely, that many gold standard protein complexes are identified with statistical significance from diverse ranges of scoring metrics—supports a commonly held practice in CF-MS data analysis: the use of panels of scoring metrics in machine learning classifiers to improve the broad-scale precision of protein complex identification (see Table I). However, Fig. 3 also shows that several commonly employed scoring metrics—Euclidean distance, NCC and Peak Location—do not generally appear to be ideal candidates for use in such panels, as they lack relative broad-scale sensitivity.

The additional findings described in relation to Fig. 3—that scoring metrics of low overall sensitivity may nonetheless have unique advantages for particular subsets of protein complexes—are an underappreciated aspect of CF-MS data analysis. If a CF-MS study is being performed with the goal of maximizing the recall of protein complexes, it may therefore be beneficial to make use of a panel of scoring metrics employed across separate stand-alone analyses (as opposed to a panel within a single machine learning classifier). When seeking to identify protein complexes whose subunits are, for example, frequently involved in multiple interactions, and thus produce fractionation profiles with multiple peaks, mutual information-based metrics may have advantages. Similarly, if seeking to identify protein complexes that only exist at specific masses, such as those without diverse sets of post-translational modifications, the Peak Location metric may have particular advantages when working with CF-MS data obtained via SEC.

Orthogonal Fractionations Are More Efficient than Stand-Alone Fractionation—Past CF-MS studies have employed a variety of fractionation methods, as outlined in Table I. These differences in experimental design have, in part, been driven by the types of protein complexes under investigation. SEC and IEX have generally been used to study soluble complexes, whereas BN-PAGE and mild detergent-based separations have been used for membrane-associated complexes (53). The varying degrees of fraction collection employed in these studies have been influenced by the resolution capable of being achieved using these different forms of separation.
Although these factors have influenced how past studies have been performed, experimental design for CF-MS data collection has nonetheless remained imprecise. One factor contributing to this imprecision has been the aforementioned lack of consensus on how many fractions should be collected to produce fractionation profiles that are optimal for CF-MS data analysis, if presented with a given chromatographic resolving power. Moreover, the relative recall versus experimental cost of single versus orthogonal separations has not yet been described.

In considering how much fraction collection should be performed, a finding of Fig. 4A – that the benefits of extensive fractionation of cell lysates may often not outweigh the costs – indicates that an assumption in the CF-MS literature may require re-interpretation. This assumption states that in order to improve the precision of CF-MS data analysis, the elution of unrelated proteins in the same fraction (i.e. chance co-fractionation) should be minimized, for example via very high resolution chromatography combined with extensive fraction collection (53). However, the present results indicate that chance co-fractionation is not inherently problematic. This is most strongly evidenced by the human SEC versus IEX CF-MS data sets. The average peak widths in the human SEC data set are higher than those in the human IEX data set (according to PrinCE; data not shown), indicating that higher resolution chromatography was achieved via IEX. Moreover, higher resolution fractionation profiles were collected for the IEX data set relative to the SEC data set (108 versus 40 fractions collected respectively). Despite this, Fig. 3A shows that after correcting for the 26% fewer fractionation profiles observed in the human IEX data set relative to the human SEC data set, recall of gold standard protein complexes in the IEX data set is not different from that in the SEC data set (54). This indicates that present implementations of these approaches do not maximize recall. Both Reference Complex Profiling and CCprofiler identify known protein complexes using fractionation profiles only. Of the fractionation profiles likely to be associated with known protein complexes shown in Fig. 5A, only limited percentages are verified as statistically significant using Reference Complex Profiling: 27% in the yeast SEC data set, 16% in the yeast IEX data set, 25% in the human SEC data set and 31% in the human IEX data set. Although it is difficult to estimate recall using these numbers, these low percentages leave open the possibility that some known protein complexes may not have reached statistical significance because their relevant fractionation profiles are not sufficiently distinctive. To remedy this without collecting additional CF-MS data, gene co-expression data, high confidence PPI data and GO may be of assistance as discussed in relation to Fig. 6.

The above approaches are inherently limited by the fact that novel protein complexes are not considered. Fig. 5A suggests that this may produce notable limitations in recall even in heavily benchmarked model organisms. Moreover, these approaches are limited to the study of organisms with existing reference libraries of protein complexes. To reach the full potential of CF-MS, identification of novel protein complexes must therefore also be considered.

To uncover both known and novel protein complexes in model organisms, it can be envisaged that the above approaches can be extended. If known protein complexes are firstly identified, novel protein complexes may then in theory be uncovered by subjecting unassigned fractionation profiles to similarity scoring. Unfortunately, genomic data integration is unlikely to add much value in such analyses, as discussed in relation to Fig. 6B.

**Analytical Guidelines for co-fractionation Mass Spectrometry**

*Mol Cell Proteomics (2020) 19(11) 1876–1895* 1891
does not preclude the use of gene co-expression data or GO to shortlist putative novel PPIs as candidates for orthogonal validation, it indicates that few, if any, novel PPIs will be uncovered in this manner.

An alternative is to characterize both known and novel protein complexes or PPIs concurrently. The supervised machine learning classifiers and clustering methods of Table I fall under this category. Unfortunately the current findings indicate that present implementations of these methods are likely to be imprecise. This is because the machine learning classifiers in Table I frequently employ features associated with external genomic data, which can have different characteristics for known and novel PPIs as discussed in relation to Fig. 6B. The common practice of using known PPIs to train and test classifiers is therefore problematic, which helps explain why novel PPIs uncovered using these classifiers may have very high false discovery rates (9). Moreover these differences preclude the use of GO to define cluster cutoff distances when employing the clustering methods of Table I. Together these insights are strongly consistent with and shed new light on the recent finding that genomic data integration decreases the power to uncover novel PPIs (52).

One possible avenue toward precise characterizations of novel protein complexes is to identify conserved protein complexes across organisms. This has, for example, been attempted using CF-MS data sets collected from multiple metazoan (5) or plant (13) species. The present comparisons between yeast and human identify a series of putative novel conserved PPIs, listed in supplemental Table S1, supporting this possibility. Consistent with the findings described in Fig. 6, these putative novel PPIs could not have been uncovered solely by shortlisting proteins with shared GO or gene co-expression. This highlights the unique means by which they were uncovered, but also their need to be validated using orthogonal experimental data.

**Extending Characterizations of Protein Complexes to NonModel Organisms**—The above discussions are centered on the analysis of CF-MS data from model organisms. Together they also provide guidance for the analysis of nonmodel organisms. It is common practice to use orthologous genomic data from model organisms — including orthologous PPIs — to interpret CF-MS data obtained from organisms without extensive reference libraries of protein complexes (9, 11, 13, 55). The findings of Fig. 5B indicate this will only be effective for certain subsets of fractionation profiles. Specifically, there is a high possibility that fractionation profiles from nonmodel organisms with characteristics that match those of putative novel interactions in model organisms cannot be assessed using orthologous genomic data. They should therefore be omitted from CF-MS data analysis as detailed in the section below.

If analyzing nonmodel organisms using fractionation profiles alone, and not orthologous genomic data, the tools available for increasing precision are further reduced. Following Reference Complex Profiling, large proportions of well correlated pairs of fractionation profiles (i.e. those making up the colored bars in Fig. 5A) cannot immediately be attributed to statistically significant protein complexes: 80 ± 7% and 76 ± 1% in the yeast and human CF-MS data sets respectively. It is likely that some of these nonsignificant pairs represent genuine interactions, as discussed earlier. Nonetheless these high percentages indicate that precise PPI identifications are difficult when using stand-alone fractionation profile similarity scoring. The practice of shortlisting putative conserved PPIs (5, 13) may therefore be particularly advantageous for nonmodel organisms. For example, if the well-matched fractionation profiles in Fig. 5A are limited to orthologous yeast and human pairs, 97 and 95% can be attributed to known yeast and human PPIs respectively. This suggests that PPIs uncovered by shortlisting conserved fractionation profile pairs may be relatively precise.

**Analytical Recommendations for co-Fractionation Mass Spectrometry**—A series of analytical guidelines and recommendations for the collection and analysis of CF-MS data, based on the insights discussed above, are presented below.

In relation to CF-MS data collection, it is possible to prioritize cost efficiency while maintaining data quality. If employing stand-alone SEC or IEX of complex cell lysates using methods like those studied here, collecting as few as 40 fractions will produce fractionation profiles with sufficient resolution for near-maximum recall of protein complexes. If collecting additional fractions to maximize recall, the benefits of spreading fractions across complementary data sets created using orthogonal separations, rather than a single higher resolution data set, should be the primary driving force in experimental design. Higher resolution stand-alone data sets may, however, be appropriate if low abundance proteins, such as those potentially associated with novel protein complexes in model organisms, are a specific focus of investigation.

In relation to data analysis, precise analyses are possible if profiling known protein complexes via, for example, Reference Complex Profiling or CCProfiler (14). If seeking to maximize the recall of such methods using a stand-alone scoring metric, any one of the following correlation metrics — Pearson, Spearman or Kendall — or the Co-apex metric is recommended. However, if additional computation time is an option, the improvements in recall that can be expected from the union of a panel of scoring metrics should be taken advantage of. Large panels will maximize recall (see supplemental Table S2); however small panels of well-chosen correlation-based, distance-based, mutual information-based or peak-based metrics will also be effective while reducing computation time. The following panel, which identifies 25.0% to 82.3% more statistically significant protein complexes from the present CF-MS data sets compared with Spearman correlations alone (supplemental Table S2), is recommended as one such option: Spearman correlation, DMax, BCMI and Co-apex. If collecting data using orthogonal fractionation methods, it is recommended that any
Precise CF-MS analyses of novel protein complexes are not yet possible when studying individual organisms, even when genomic data integration and the collection of high-resolution CF-MS data are considered. This holds true for both model and nonmodel organisms. CF-MS studies of novel protein complexes and PPIs in individual organisms should therefore be limited to generating shortlists for targeted orthogonal experimental validation. To maximize the quality of these shortlists, false discovery rates can be reduced via the following recommendations. A) If training and testing a machine learning classifier using gold standard protein complexes from a model organism, the classifier should not be used to identify novel protein complexes in this organism if it employs features derived from genomic data. B) If studying a nonmodel organism, machine learning classifiers employing features derived from orthologous genomic data can be used. However, if training the classifier using orthologous gold standard protein complexes, it is recommended that the following fractionation profiles should not be analyzed using the classifier: the lowest abundance quartile, as defined using maximum peak heights; those with peaks in the last quartile of retention times if employing SEC; and those with peaks in the first quartile of retention times if employing mixed-bed IEX. C) If predicting novel protein complexes using stand-alone fractionation profile similarity scoring, Pearson, Spearman or Kendall correlations or the Co-apex metric are recommended, whereas Euclidean distance is not. Moreover, in this context, panels of scoring metrics are generally not recommended outside of a machine learning framework. This is because the intersection of results derived from such panels may result in high false discovery, because precision cannot be accurately measured, whereas the intersection of results will often lead to very low recall (see supplemental Figs. S9–S12). D) If employing clustering methods to define the subunit compositions of novel protein complexes, the use of external genomic data to define cluster cutoff distances is not recommended.

Finally, the study of CF-MS data sets from multiple organisms presents additional options for identifying novel protein complexes. The present findings lend support to the practice of identifying putative conserved PPIs across organisms to improve precision.

CONCLUSIONS

The substantial promise of CF-MS lies not only in its potential for very broad-scale characterizations of endogenous and unmanipulated protein complexes, but also its inherent capacity to study how these protein complexes change (17). Recent advances in CF-MS have brought the method closer to fulfilling this potential (7, 15). However, a primary objective of such work—to enable routine time course or cohort analyses of cellular assemblies (15)—is still experimentally prohibitive, and requires increasingly effective CF-MS data analysis workflows to be developed.

The present recommendations for more cost-effective CF-MS data collection will bring this objective closer. Moreover, the present recommendations for precise and sensitive CF-MS data analysis, when integrated with workflows such as CCprofi ler (14, 15), should further fulfill the data analysis requirements of this objective for known protein complexes in model organisms. Together these advances can be expected to bring routine and impactful studies of the dynamics of known PPIs well within the reach of CF-MS.

The present findings also highlight the pronounced differences in studying known versus novel protein complexes using CF-MS. It can therefore be envisaged that CF-MS will immensely benefit from increasingly sophisticated reference libraries of protein complexes and PPIs. Targeted orthogonal experimental validation of putative novel complexes shortlisted via CF-MS may be an effective means of expanding these reference libraries. If following the above recommendations, high quality shortlists in nonmodel organisms will comprise mainly of large and high abundance protein complexes, whereas they will contain more small and low abundance protein complexes in highly studied model organisms. These differences suggest that validation experiments can be precisely targeted for different organisms. For example, high depth of coverage chemical cross-linking MS experiments performed on SEC fractions corresponding to selected molecular weight ranges can be envisaged. When coupled with proteomics and CF-MS data sharing initiatives (5, 16, 56), expanding reference libraries of protein complexes in this manner will pave a way forward for CF-MS studies of continually increasing depth, precision, and impact.

DATA AVAILABILITY

The MS proteomics data have been deposited to the Proteomexchange Consortium via the PRIDE (57) partner repository with the data set identifier PXD019513.

Acknowledgments—We thank Dr. Ling Zhong, Ms. Sydney Liu Lau and A/Prof. Mark Raftery for their maintenance of the mass spectrometers housed at the UNSW Bioanalytical Mass Spectrometry facility. This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney. We thank Dr. Michele Tinti for his generous advice, and Prof. Paul Haynes for his critical evaluations of manuscript drafts.

Funding and additional information—G.H.-S. and M.R.W. thank the Australian Research Council (ARC) for financial
support. G.H.-S. also acknowledges funding from the UNSW School of Biotechnology and Biomolecular Sciences.

**Author contributions**—C.N.I.P., S.B., J.J.H., and G.H.-S. designed research; C.N.I.P., S.B., D.W., L.M.T., J.J.H., and G.H.-S. performed research; C.N.I.P., S.B., D.W., J.J.H., J.G., and G.H.-S. contributed new reagents/analytic tools; C.N.I.P., S.B., D.W., J.J.H., M.R.W., and G.H.-S. analyzed data; C.N.I.P., S.B., J.J.H., and G.H.-S. wrote the paper.

**Conflict of interest**—The authors declare that they have no conflicts of interest with the contents of this article.

**Abbreviations**—The abbreviations used are: AGC, automatic gain control; AP-MS, affinity purification-mass spectrometry; AUROC, area under the receiver operating characteristic curve; BCMI, Biased Corrected Mutual Information; BN-PAGE, blue native polyacrylamide gel electrophoresis; CF-MS, co-fractionation mass spectrometry; CORUM, the comprehensive resource of mammalian protein complexes; DDA, data-dependent acquisition; DMax, maximum cumulative distance between two scaled vectors; EGAD, Extending ‘Guilt-by-Association’ by Degree; GMIC, Generalized Mean Information Coefficient; GO, Gene Ontology; HCD, higher-energy collision dissociation; IEX, ion exchange chromatography; MI, Mutual Information; M115, 057117; M115. 057117

REFERENCES

1. Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011) Interactome networks and human disease. Cell 144, 986–998

2. Ideker, T., and Krogan, N. J. (2012) Differential network biology. Mol. Syst. Biol. 8, 565

3. Bonetta, L. (2010) Protein–protein interactions: interactome under construction. Nature 468, 851–854

4. Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Feng, Y., Phanse, S., Babu, M., Craig, S. A., Hu, P., Wan, C., Vlasblom, J., Dar, V-U-N., Bezginov, A., Clark, G. W., Wu, G. C., Wodak, S. J., Tillier, E. R. M., Paccanaro, A., Marcotte, E. M., and Emilie, A. (2012) A census of human soluble protein complexes. Cell 150, 1088–1098

5. Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A., Chessman, K., Pal, S., Croman, G., Papoulas, O., Ni, Z., Boutz, D. R., Stoilova, S., Havugimana, P. C., Guo, X., Malty, R. H., Sarov, M., Greenblatt, J., Babu, M., Derry, W. B., Tillier, E. R., Wallfording, J. B., Parkinson, J., Marcotte, E. M., and Emilie, A. (2015) Panorama of ancient metazoan macromolecular complexes. Nature 525, 339–344

6. Drew, K., Müller, C. L., Bonneau, R., and Marcotte, E. M. (2017) Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. PLoS Comput. Biol. 13, e1005625

7. Scott, N. E., Rogers, L. D., Prudova, A., Brown, N. F., Fortelny, N., Overall, C. M., and Foster, L. J. (2017) Interactome disassembly during apoptosis occurs independent of caspase cleavage. Mol. Syst. Biol. 13, 906

8. Larance, M., Kirkwood, K. J., Tinti, M., Murillo, A. B., Ferguson, M. A., and Lamond, A. I. (2016) Global membrane protein interactome analysis using in vivo crosslinking and MS-based protein correlation profiling. Mol. Cell. Proteomics 0115, 055467

9. Shtatsky, M., Dong, M., Liu, H., Yang, L. L., Choi, M., Singer, M. E., Geller, J. T., Fisher, S. J., Hall, S. C., and Hazen, T. C. (2016) Quantitative tagless co-purification: a method to validate and identify protein-protein interactions. Mol. Cell. Proteomics 115, 057117

10. Stacey, R. G., Skinnider, M. A., Scott, N. E., and Foster, L. J. (2017) A rapid and accurate approach for prediction of interactomes from co-elution data (PPrICE). BMC Bioinformatics. 18, 457

11. Crozier, T. W., Tinti, M., Larance, M., Lamond, A. I., and Ferguson, M. A. (2017) Prediction of protein complexes in Trypanosoma brucei by protein correlation profiling mass spectrometry and machine learning. Mol. Cell. Proteomics 0117, 068122

12. Carlson, M. L., Stacey, R. G., Young, J. W., Wason, I. S., Zhao, Z., Rattray, D. G., Scott, N., Kerr, C. H., Babu, M., Foster, L. J., and Duong Van Hoa, F. (2019) Profiling the Escherichia coli membrane protein interactome captured in Peptidus libraries. Elife 6,

13. McWhire, C. D., Papoulas, O., Drew, K., Cox, R. M., June, V., Dong, O. X., Kwon, T., Wan, C., Saimi, M. L., Roux, S. J., Browning, K. S., Chen, Z. J., Ronald, P. C., and Marcotte, E. M. (2020) A pan-plant protein complex map reveals deep conservation and novel assemblies. Cell 181, 460–474, e14

14. Heusel, M., Bludau, I., Rosenberger, G., Hafen, R., Frank, M., Banaei-Esfahani, A., Drogen, A., Collins, B. C., Gstaiger, M., and Aebersold, R. (2019) Complex-centric proteome profiling by SEC-SWATH-MS. Mol. Syst. Biol. 15, e8435

15. Heusel, M., Frank, M., Köhler, M., Amon, S., Frommelt, F., Rosenberger, G., Bludau, I., Aulakh, S. K., Linder, M. I., and Liu, Y. (2019) A global screen for assembly state changes of the mitotic proteome by SEC-SWATH-MS. CELL-SYSTEMS-D-19-00261

16. Kirkwood, K. J., Ahmad, Y., Larance, M., and Lamond, A. I. (2013) Characterisation of native protein complexes and protein isoform variation using size-fractionation based quantitative proteomics. Mol. Cell. Proteomics

17. Kristensen, A. R., Gasperon, J., and Foster, L. J. (2012) A high-throughput approach for measuring temporal changes in the interactome. Nat. Methods 9, 907–909

18. Scott, N. E., Brown, L. M., Kristensen, A. R., and Foster, L. J. (2015) Development of a computational framework for the analysis of protein correlation profiling and spatial proteomics experiments. J. Proteomics 118, 112–129

19. Stacey, R. G., Skinnider, M. A., Chik, J. H., and Foster, L. J. (2018) Context-specific interactions in literature-curated protein interaction databases. BMC Genomics 19, 758

20. Cusick, M. E., Kitgord, N., Vidal, M., and Hill, D. E. (2005) Interactome: gateways into systems biology. Human Mol. Gen. 14, R171–R181

21. Benschop, J. J., Brabers, N., van Leenen, D., Bakker, L. V., van Deutkom, H. W., van Berkum, N. L., Apweiler, E., Lijnzaad, P., Holstege, F. C., and Kemmeren, P. (2010) A consensus of core protein complex compositions for Saccharomyces cerevisiae. Mol. Cell 38, 916–928

22. Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B. D. M., Burston, H. E., Viveacoumar, F. J., Snider, J., Phanse, S., Feng, V., Tam, Y. Y. C., Davey, M., Hnatshak, O., Bajaj, N., Chandran, S., Punna, T., Christopoulos, C., Wong, V., Yu, A., Zhong, G., Li, J., Staglier, J., Conibeer, E., Wodak, S. J., Emilie, A., and Greenblatt, J. F. (2012) Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. Nature 489, 585–589

23. Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, H., Hudak, M., Michon, A.-M., Schelder, M., Schrie, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rich, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) A pan-plant protein complex map reveals deep conservation and novel assemblies. Nature 440, 631–636

24. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tiskakis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Men, A., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Plistone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altfat-
