A machine learning-based approach to identify reliable gold standards for protein complex composition prediction

Pengcheng Yang[1,‡], Youngwoo Lee[2,3,‡,*], Daniel B Szymanski[2,3,4], and Jun Xie[1,*]

[1] Department of Statistics, Purdue University, West Lafayette, Indiana, 47907, USA

[2] Department of Botany and Plant Pathology, Purdue University, West Lafayette, Indiana, 47907, USA

[3] Center for Plant Biology, Purdue University, West Lafayette, Indiana, 47907, USA

[4] Department of Biological Sciences, Purdue University, West Lafayette, Indiana, 47907, USA

[*] Corresponding authors: Youngwoo Lee (lee1932@purdue.edu) and Jun Xie (junxie@purdue.edu)

[‡] These authors contributed equally to this work.

Abbreviations

CFMS: co-fractionation mass spectrometry

CORUM: the comprehensive resource of mammalian protein complexes database

SEC: size exclusion chromatography

M_{app}: protein apparent mass

M_{mono}: protein monomeric mass

M_{calc}: protein calculated mass

R_{app}: multimerization state or determinator

WCC: weighted cross-correlation

Euclid: Euclidean distance

d: distance between two proteins

SOM: self-organizing map

AP: affinity propagation

Abstract

Co-Fractionation Mass Spectrometry (CFMS) enables the discovery of protein complexes and the systems-level analyses of multimer dynamics that facilitate responses to environmental and developmental conditions. A major challenge in the CFMS analyses, and other omics approaches in general, is to conduct validation experiments at scale and develop precise methods to evaluate the performance of the analyses. For protein complex composition predictions, CORUM is commonly used as a source of known complexes; however, the subunit pools in cell extracts are very rarely in the assumed fully assembled states. Therefore, a fundamental conflict

exists between the assumed multimerization of the CORUM "gold standards" and the CFMS experimental datasets to be evaluated. In this paper, we develop a machine learning-based "small world" data analysis method. This method uses size exclusion chromatography profiles of predicted CORUM complex subunits to identify relatively rare instances of fully assembled complexes, as well as bona fide stable CORUM subcomplexes. Our method involves a two-stage machine learning approach that integrates information from CORUM and CFMS experiments to generate reliable gold standards of protein complexes. The predictions are evaluated by both statistical significance and size comparison between calculated and predicted complexes. These validated gold standards are then used to assess the overall reliability of CFMS-based protein complex composition predictions.

Introduction

Co-fractionation mass spectrometry (CFMS) is a high-throughput, mass spectrometry-based protein quantification coupled with biochemical fractionation methods to analyze protein complex compositions under non-denaturing conditions. This "guilt by association" method was initially used to predict protein organelle localization based on co-elution with known marker proteins (1) and was subsequently applied to predict protein interactors (2, 3). The technique has evolved from the principle that proteins present in a stable complex co-migrate independent of the separation method used. In plant systems, CFMS has been extensively employed for the determination of apparent masses, localization, and compositions of protein complexes across a wide variety of species and tissue types, including leaves, roots, and flowers (4-9), as well as in organelles like chloroplasts (10) and mitochondria (11, 12). CFMS has broad use as a valuable

tool to analyze protein complex dynamics, including circadian changes (13), protein-ligand interactions (14, 15), and multimerization variants across plant species (16). A remaining challenge is to determine the appropriate data types and profile analysis methods to generate the most accurate protein complex composition predictions (17, 18). A major impediment to progress in this area is the lack of a large set of reliable gold standards to evaluate prediction accuracies.

Protein complex prediction methods vary wildly among different studies. Many are achieved by using multiple metrics from external data sources that could inform multimerization behaviors. Examples include using mRNA co-expression or co-citation information via machine learning classifiers (3, 9, 18-20) or integration of existing protein interaction predictions from orthogonal approaches (4-6, 16, 21-24). There is no clear-cut agreement on the effectiveness of those strategies for protein complex predictions (18). All approaches rely on known protein complexes as gold standards from a reference database, like the CORUM mammalian protein complex database (25). Gold standard protein complex datasets are critical because they define accuracy measures, e.g., the precision and recall, of protein complex predictions. Inaccurate gold standards result in a wrong validation dataset, misleading the prediction model to unreliable predictions.

The CORUM protein complexes are widely used as assumed gold standards to evaluate protein complex predictions. During the prediction evaluation, pairs of subunits in the reference CORUM complexes comprise a positive set of protein-protein interactions. Negative interactions are created from proteins not present in the positive interaction set. These positive and negative sets have been used for training and testing computation methods of protein complex predictions (9, 18, 20, 26). This approach assumes that CORUM complexes are fully assembled in the cell

extracts that are used as the input for CFMS experiments. However, in our previous CFMS experiments (6, 16, 22, 23), many CORUM complex subunits have an apparent mass similar to their theoretical monomeric mass. Furthermore, in several recent publications, weak correlations were reported among subunits of CORUM complexes in the CFMS datasets (6, 16). Pang et al. (18) displayed probability density plots of the Pearson correlations between fractionation profiles of shared subunits of CORUM human complexes (Figure 2 in (18)), in which the density plots have a mode around 0 correlation values. A similarly low correlation among subunits of CORUM complexes was reported in Extended Data Figure 2 of Wan et al. (20). These plots and our experimental data indicate that the most abundant cellular pools of CORUM complex subunits, which are what are primarily detected in LC/MS, are not necessarily in the fully assembled state. In other words, subunits of orthologous proteins from a CORUM complex do not always co-elute, suggesting that CORUM complexes are unlikely to be reliable gold standards for protein complex predictions.

In this paper, we have developed a machine learning method to identify reliable CORUM complexes or subcomplexes, which can then be used as gold standards for CFMS analysis. More specifically, we collect plant proteins that are orthologous to subunits of a CORUM complex. Next, we conduct a “small world” analysis, examining the CFMS elution profile patterns of every orthocomplex one by one. Subcomplex predictions are generated using a robust unsupervised machine learning method, namely self-organizing map (SOM). We use a limited set of well-known rice complexes to train the algorithm, taking advantage of its unsupervised nature. The developed method is proficient in identifying reliable complexes with similar CFMS elution profiles. Statistical significance is calculated for the predictions using Monte Carlo simulations. A gold standard is defined as a CORUM complex or subcomplex that is composed

of subunits with analogous CFMS elution profiles, is statistically significant, and has consistent apparent mass and calculated mass. We further applied our validated gold standards and the metrics of intactness and purity to demonstrate the method's potential for evaluating protein complex composition predictions.

Materials and Methods

External SEC data acquisition and filtering

In this study, we used a published rice CFMS dataset, which was previously used for protein complex predictions. The data were downloaded from the Supplemental Data Sets S2 and S3 (16). The dataset contains two replicates of size exclusion chromatography (SEC) fraction profiles and protein apparent mass (M_{app}: measured mass from SEC experiments), monomeric mass (M_{mono}), and multimerization state (R_{app}: multimerization determinator, which is defined as a ratio of M_{app} to M_{mono}). We filtered the data as described below to identify reliable protein profiles for the gold standard complex discovery. Reproducible protein profiles were selected if their elution peak locations between the two replicates were within 2-SEC fractions. If the reproducible peak was located at the first (void) fraction, it was defined as an unresolvable peak and removed from the dataset. When a protein profile had more than one reproducible peak, we separated the distinct peaks and split the data into multiple segments one by one peak. For a peak separated from a multiple-peak protein, it was treated as an individual profile in the following data analysis if its mean M_{app} from the two replicates was less than 850 kDa. They were annotated with a numerical suffix appended to the protein name to indicate peak numbers for each of the two replicates in the Supplemental Tables.

Interkingdom ortholog mapping for CORUM orthocomplex assignment

To infer human-to-rice orthologs, human proteome search database (9606.fasta) was obtained from InParanoid 8 website (https://inparanoidb.sbc.su.se/), and rice proteome (Osativa_323_v7.0.protein_primaryTranscriptOnly.fa) was downloaded from Phytozome V12 database (https://phytozome-next.jgi.doe.gov/). The InParanoid software Version 4.1 was used to infer orthologs between human and rice species (27). The inferred orthologs were reported in Supplemental Table S1. To identify rice orthologous complexes (orthocomplexes) corresponding to CORUM complexes, human protein complexes were downloaded from the CORUM database (http://mips.helmholtz-muenchen.de/corum/). According to the ortholog groups specified in Supplemental Table S1, the subunits in the human protein complexes were converted to their corresponding rice orthologs (Supplemental Table S2). A plant species has experienced polyploidization and gene duplication, creating different levels of genetic redundancies across species (28). This genome-wide complexity makes the ortholog assignment challenging. It is common that multiple rice orthologs/paralogs were mapped to a single human ortholog. In such cases, we treated all rice paralogs as members of a rice orthocomplex but considered them a single “ortho-paralog” group to calculate the rice proteome overlap with the human CORUM complexes. The subunit coverage of a rice orthocomplex is defined as the ratio of the number of subunits with inferred orthologs to the total number of subunits of a CORUM complex. Those complexes with a subunit coverage greater than 2/3 were chosen as high-coverage rice orthocomplexes for gold standard complex predictions.

Integration of rice orthocomplexes into rice SEC data

After assigning rice orthocomplexes through the mapping of human CORUM complexes,

we integrated this orthocomplex information with the experimental SEC profile data. We identified useful rice orthocomplexes that contained at least two subunits in the SEC profile data. Subsequently, we refined these orthocomplexes by eliminating redundant orthocomplexes consisting of identical rice subunits, smaller orthocomplexes that were subsets of larger ones, and orthocomplexes comprising solely rice subunits mapped to a single human ortholog.

Experimental Design and Statistical Rationale

Distance metric

In the small-world analysis, we utilized a distance metric that combines two similarity measurements of a pair of proteins in an orthocomplex, i.e., the correlation of the fractionation profiles between the pair and the distance of peak fraction locations between the pair. First, a weighted cross-correlation (*WCC*) between a pair of protein profiles was calculated by the following formula (29)

$$WCC(i,j) = \frac{x_i^T W x_j}{\sqrt{x_i^T W x_i}\sqrt{x_j^T W x_j}} \qquad (1).$$

Here, x_i and x_j were two column vectors representing the SEC profiles of the two proteins in the orthocomplex. The weight matrix, denoted as *W*, was constructed with ones on the main diagonal, but elements on the sub-diagonal and the super-diagonal decreased proportionally as they moved away from the main diagonal. All elements outside of the sub- and super-diagonals were set to zero. We set the bandwidth of *W* to be 2 and used a weighting function of (1 − distance/3). That is, there were 2 sub-diagonals and 2 super-diagonals in the weight matrix *W*, and the weights on the first sub/super-diagonals were 2/3, while the weights on the second sub/super-diagonals were 1/3. This choice of parameters enables *WCC* to extract the similarity information across the peak profiles within a neighborhood of 2 fractions. A distance (*WCCd*)

based on the weighted cross-correlation for a pair of proteins was given by

$$WCCd(i,j) = 1 - WCC(i,j) \qquad (2).$$

Another distance metric between the pair of proteins was calculated from their corresponding peak locations as follows:

$$Euclid(i,j) = |y_i - y_j| \qquad (3)$$

where y_i and y_j were the peak fraction locations of a pair of proteins, obtained using the Gaussian peak fitting algorithm (6). These peak locations had been standardized by dividing them by the largest peak location of a complex, ensuring they fall within the range of 0 to 1 for the Euclidean distance calculation in Formula (3). Both distance measurements were computed for the two replicates of the SEC data. An overall distance between two proteins was defined as

$$d_{overall} = w \times (WCCd_1 + WCCd_2) + (0.5 - w) \times (Euclid_1 + Euclid_2) \qquad (4)$$

where $WCCd_1$ and $WCCd_2$ were the weighted cross-correlation distance, and $Euclid_1$ and $Euclid_2$ were the Euclidean distance between the peak locations of the pair of proteins. There were two distances indicated by the subscripts 1 and 2 for the two SEC data replicates. The combination weight w was obtained by model training described in the following.

Pre-clustering evaluation of orthocomplexes

Before conducting an unsupervised clustering analysis as described next, we first assessed whether an orthocomplex formed a single cluster. This step also helped us evaluate which orthocomplexes were potentially fully assembled CORUM complexes, as a fully assembled orthocomplex must be a single cluster. More specifically, in cases where an orthocomplex consisted of only two subunits in the SEC profile data, these small orthocomplexes were excluded from the subsequent clustering analysis and were labeled as “only 2 proteins in the complex before clustering” in the result tables. However, as later defined, these instances

were subjected to a bootstrap *p*-value calculation to determine if the two subunits had the potential to form a complex. If the resulting *p*-value fell below the 5% threshold, the two subunits had a significantly small distance and hence were predicted to form a single cluster.

For orthocomplexes with at least three subunits from the SEC data, we assessed whether all subunits exhibited similar profiles, which was indicated by significantly small pairwise distance metrics. Two distance metrics, defined in Formula (2) and Formula (3), were utilized for this evaluation. If the distances calculated for all subunit pairs within an orthocomplex fell below the 5th percentile threshold of an empirical distribution, we inferred that all subunits within that orthocomplex were part of the same complex. The empirical distribution was constructed by calculating pairwise distances among all proteins within the SEC data. We systematically examined both distance metrics, and any orthocomplexes that met the 5th percentile criterion for either distance metric were predicted to form a single cluster. In Supplemental Table S3, column T, those meeting the *WCCd* distance criterion were labeled as “all proteins defined in the complex by similarity before clustering”, while those meeting the Euclidean distance criterion were labeled as “all proteins defined in the complex by Gaussian peak distance before clustering”.

Note that some of these orthocomplexes were potentially fully assembled CORUM orthologs, as our analysis indicated they formed a unified complex that could not be separated. To rigorously confirm the full assembly, we further examined whether these inseparable orthocomplexes consisted of all CORUM subunits as well as if the apparent mass observed in the SEC experiment agreed with the calculated mass summing over all subunits (more details in the size evaluation section below). Our confirmation of orthocomplexes as fully assembled was substantiated by satisfying these additional two criteria.

Two-stage clustering algorithm for subcomplex prediction

Only a very small number of rice orthocomplexes were found to be fully assembled. For the majority of the orthocomplexes, we conducted a small-world clustering analysis on each one individually. The small-world analysis employed a two-stage clustering algorithm with the distance metric defined in Formula (4). Our two-stage procedure is composed of a clustering step by self-organizing map (SOM) algorithm (30) and a cluster merging step by the affinity propagation (AP) algorithm (31). SOM is a machine learning method for clustering analysis. It is a specific type of neural network model that is designed to represent SEC profiles using robust features. As a result, the method exhibits resistance to random noises and possesses the advantage of robustness against data measurement errors or missing values. Employing the SOM algorithm, the subunits of a rice orthocomplex were clustered, leading to the formation of distinct subgroups. Importantly, these distinct subgroups were predicted to represent subcomplexes within the larger orthocomplex.

To yield the optimal final clusters, a relatively large number of clusters was initially selected and then followed by the merging of resulting clusters using the AP algorithm. For this two-step clustering analysis, we fine-tuned three parameters of the clustering algorithms, including the weight w on $d_{overall}$ in Formula (4), the number of clusters in SOM, and the merging threshold for the AP algorithm. To determine these parameters, we used four well-known rice complexes: 19S proteasome, 20S proteasome, 14-3-3 hetero/homooligomers, and the exosome. We determined the three parameters of our algorithm so that subunits for each of these known complexes were clustered together. These known complexes are part of four CORUM orthocomplexes with the following IDs: PA700−20S−PA28 complex (CORUM ID: 193), RAF1−MAP2K1−YWHAE complex (CORUM ID: 5873), HSF1−YWHAE complex (CORUM

ID: 2145), and Exosome (CORUM ID: 789), as indicated in Supplemental Table S3.

The trained two-stage clustering algorithm was implemented across the rice orthocomplexes one by one. For each rice orthocomplex, the two-stage clustering approach could yield clusters with multiple members or a single member. Clusters with multiple members were designated as subcomplexes, while single-member clusters were labelled as “singletons”. In cases where all members of a predicted subcomplex were mapped to a single human ortholog, it was called an “ortho-paralog subcomplex”. Altogether, the results were labeled as “subcomplex i”, “subcomplex (ortho-paralog) i”, and “singleton i”, where i is a subcomplex identification, in the Supplemental Tables and Figures.

Furthermore, based on the external input data, proteins with small R_{app} values, specifically $R_{app} \leq 1.6$, were considered as putative monomers. When our algorithm identified proteins with small R_{app} values as singletons, we deemed these monomer predictions accurate. This is because proteins with the small range of R_{app} were anticipated to be monomers, and the clustering algorithm affirmed this anticipation.

Statistical significance and size evaluation of identified subcomplexes

The subcomplexes, as well as the potentially fully assembled complexes, identified through the small-world analysis, were evaluated using statistical *p*-values and a comparison of apparent mass (M_{app}) and calculated mass (M_{calc}) to identify the gold standards (Figure 1). First, a bootstrap *p*-value for a subcomplex/complex was calculated using Monte Carlo simulations. The mean distance for pairs of proteins in a subcomplex was calculated using the overall distance, $d_{overall}$, in Formula (4). A random complex was generated by randomly sampling the same number of proteins in the identified subcomplex from all SEC profiles in the CORUM

orthocomplex profile dataset. Protein subunits in the random complex were not related, hence their distance $d_{overall}$ followed a pure random distribution. The mean distance for pairs of proteins present in the random subcomplex was calculated. The random sampling was repeated 134,350 times, equal to 50 times the number of the rice SEC profiles. The *p*-value was calculated as the fraction of times the random mean distance was smaller than the observed mean distance of the identified subcomplex.

For the M_{app} and M_{calc} comparison, the M_{calc} of an identified subcomplex/complex was obtained by summing the monomeric mass values of all subunits in the given complex. For an orthocomplex that contained multiple rice paralogs mapping to a single human ortholog, their M_{mono} values were averaged in calculating the complex M_{calc}. Notably, information regarding the stoichiometry of subunits within a complex is rarely available (25). Thus, we manually curated the most common stoichiometry information in the RCSB Protein Data Bank (RCSB PDB) for the selected rice orthocomplexes (Supplemental Table S3, columns X and Y) and incorporated this stoichiometry information when calculating M_{calc}.

For each subcomplex/complex with *p*-value < 0.05, its M_{app} was compared with its calculated mass M_{calc} (Supplemental Table S4). Identified subcomplexes/complexes with similar M_{app} and M_{calc} were deemed gold standards. More specifically, M_{app} for a subcomplex/complex was given by

$$M_{app} = \frac{\sum_{i=1}^{n} {M_{app}}^{i}}{n} \qquad (5)$$

where *n* was the number of subunits in the subcomplex/complex, and ${M_{\mathrm{app}}}^{i}$ was the apparent mass of a subunit. Let M_{calc} denote the sum of the monomeric masses of subunits within the identified subcomplex/complex. We determined a subcomplex/complex with *p*-value < 0.05 as a gold standard if it satisfied the following criterion:

$$\frac{1}{2}M_{calc} < M_{app} < 2M_{calc} \qquad (6).$$

This criterion signified that the relative difference between the SEC experiment mass (M_{app}) and the mass calculated from the identified subcomplex/complex subunit composition (M_{calc}) was constrained within a factor of 1. In other words, subcomplexes/complexes with p-value < 0.05 that exhibited matching M_{app} and M_{calc} values, as defined by this criterion, were selected as gold standards. Notably, some of these gold standards originated from the pre-clustering complexes. When their M_{app} and M_{calc} values matched, they were confirmed as fully assembled CORUM orthocomplexes.

Use of gold standards to evaluate de novo complex predictions

We used the selected gold standards to evaluate protein complex predictions. Typical evaluations in the literature were merely based on positive and negative sets of protein-protein interactions (PPI). The use of false positive and false negative, or equivalently precision and recall, to assess CFMS data is limited conceptually to the inference of PPIs. We consider this practice inadequate because the information of PPI is rooted in pairwise interactions, while the information of a protein complex should be from a group of subunits. In other words, the evaluation of complex prediction should be grounded in the composition of the complex, a group of multiple subunits, not pairwise links.

To overcome the limitation with the use of precision and recall, we defined two complex-centric metrics: intactness and purity. Intactness and purity were used to measure the similarity between two sets of protein complexes, one set of gold standards and the other set of *de novo* predictions. Let C denote a list of de novo protein complex predictions,

$$C = \{c_1, c_2, \dots, c_n\}$$

where c_i is a predicted complex with multiple subunits. Analogously, let G denote our list of gold standards,

$$G = \{g_1, g_2, \dots, g_m\}.$$

Without loss of generality, we assume G is a subset of C, given that gold standards typically comprise a smaller set compared to the set of proteins to be predicted. In instances where the gold standard set is not entirely encompassed within the larger set of proteins for complex predictions, we will employ $G \cap C$ as the gold standard set. For each gold standard protein complex, $g_i \in G$, we define the intactness and purity as

$$\text{intactness}_{g_i} = \frac{\max_{c_j \in C} |c_j \cap g_i|}{|g_i|}, \quad \text{purity}_{g_i} = \frac{\max_{c_j \in C} |c_j \cap g_i|}{|c_k|} \qquad (7)$$

where c_k is the protein complex in C with the most overlaps with g_i, and | | denotes the number of proteins of the corresponding set. Note that when the maximum overlap between a gold standard complex and the predicted complexes, $\max_{c_j \in C} |c_j \cap g_i|$, involves only one subunit, the calculation of intactness is not meaningful. In such cases, the corresponding gold standard is completely split into different predicted complexes; thus, the intactness values are denoted as "Not Defined" (as in Supplemental Table S5).

The intactness metric quantifies how well the predicted complex captures the entirety of the proteins that are present in the gold standard complex. A high intactness value for a specific gold standard complex indicates that the predicted complex adequately represents the entire protein composition of the gold standard. In contrast, a lower intactness value suggests that the predicted complex might only partially capture the proteins within the gold standard complex. The purity metric assesses the specificity of a predicted protein complex with respect to a particular gold standard complex. It evaluates the proportion of proteins within the predicted complex that are also part of the gold standard complex. A higher purity value for a specific gold

standard complex indicates that the predicted complex primarily comprises proteins characteristic of that gold standard. Conversely, a lower purity value implies that the predicted complex contains proteins beyond the gold standard complex.

The size of the gold standard complexes affects both the intactness and purity metrics. Larger gold standards require more accurate predictions to achieve high intactness values, while smaller gold standards would attain high intactness values more easily. Conversely, larger gold standards could lead to higher purity values due to the potential for greater protein overlaps. These two metrics collectively offer an evaluation of the accuracy of protein complex predictions.

Dimerization prediction by AlphaFold Multimer on COSMIC^2

To run the AlphaFold Multimer software package v2.2.0 (32), the COSMIC^2 cloud platform was used (33). Protein sequences for each dimeric subcomplex were obtained from the rice proteome file *Osativa_323_v7.0.protein.fa* (34) in Phytozome V12 (35), and then searched against the full database (full_dbs) as default. Ranking confidence scores, a weighted combination of interface predicted Template Modeling score (ipTM) and predicted Template Modeling score (pTM), were used as model confidence metrics. The averaged model confidence score of the top 5 predicted models was used to evaluate predicted dimeric subcomplexes.

Statistical tests and data analysis

Statistical analysis was performed using R version 4.2.0 (36) on RStudio 2022.07.1 (37). The Flexible Self-Organizing Maps in Kohonen 3.0 package for R (38) and the APCluster package for R (39) were implemented for the SOM and AP algorithms, respectively. Gaussian

fitting code (https://github.com/dlchenstat/Gaussian-fitting) was run on MATLAB (R2022a). Microsoft Excel on Office 365 for Mac was used to organize and display the analyzed data.

Data and materials availability

The source code and sample input data for the small-world analysis are publicly available on GitHub (https://github.com/yangpengchengstat/R-code-S4_Class-protein-clustering-based-on-data-integration-of-corum-and-inparanoid.git). The package at the github link contains comprehensive information on running the code, description of the input data, and steps of performing hyperparameter tuning.

Results and Discussion

Figure 1 presents an overview of the workflow to identify true gold standards by integrating the CORUM database with the SEC profile data from rice tissue extracts. The overall scheme was to identify rice orthocomplex subunits based on subunit overlap in the plant and animal kingdoms and sequence similarity between the individual complex subunits. The SEC profile data from rice were mined for reproducible M_{app} measurements for the orthocomplex subunits, and statistical clustering methods were developed to group similar protein profiles from the SEC data in order to identify fully- and partially-assembled complexes in the rice cell extract. The predicted subcomplexes/complexes were further evaluated using statistical significance and size comparison between calculated mass and measured apparent mass. The validated subcomplexes/complexes were defined as gold standards. These gold standards serve as reliable references for the evaluation and analysis of future protein complex predictions.

Generating rice orthocomplexes from CORUM

To specifically analyze known complex and construct a reference library of putative gold standards genes, we identified CORUM orthologs in rice using the InParanoid algorithm (27). InParanoid search compared 20,834 human protein sequences with 42,160 proteins in rice. The algorithm assigned 5,363 human proteins and 8,178 rice proteins into 3,131 distinct orthologous groups (Supplemental Table S1). The higher number of orthologs from rice was due to an elevated gene copy number compared to humans (28). We next created predicted rice orthocomplexes from CORUM human complex compositions (25) using the ortholog dataset generated above. As discussed in the ortholog mapping section of the Methods and Materials, there were complicated scenarios when we assigned orthologs between human and rice. Frequently, multiple rice orthologs/paralogs were mapped to a single human ortholog, and vice versa. We constructed a CORUM orthocomplex by including all rice paralogs that were orthologous to each human subunit in the corresponding CORUM complex. Among the 3,047 human complexes curated in CORUM, 1,964 had at least one subunit orthologous to one or more rice proteins, and 436 of the 1,964 rice orthocomplexes appeared to be completely conserved (Figure 2A; Supplemental Figure S1; Supplemental Table S2). About half of the 1,964 rice orthocomplexes were highly conserved based on subunit coverages with more than 2 out of 3 orthologous subunits retained in the rice lineage.

To test for variability in the sizes of the rice and human orthologs, we compared predicted protein complexes between human and rice orthocomplexes that shared 100% subunit coverage. Using the summed monomeric masses of the plant and animal complex subunits, the M_{calc} distribution showed a high correlation (r = 0.986), indicating that most complex subunits possess similar complex masses (Figure 2B). Among those 436 orthocomplexes, 258 had detected subunit(s) in the SEC datasets. When M_{calc} values of these highly conserved rice

orthocomplexes were compared to their subunit M_{app} values measured in the rice SEC experiments, no correlation was found (Figure 2C). The low correlation shown in Figure 2C was not due to differences in M_{mono} of subunits in these 258 orthocomplexes, as we found a strong positive correlation between M_{mono} values for this subset of human and rice orthocomplex subunits (Figure 2D). Furthermore, 25 proteins from the rice orthocomplexes with 100% subunit coverage had $R_{app} \leq 1.6$ and hence were considered as likely monomers. The existence of monomeric subunit pools and partially assembled complexes can explain the large number of data points falling well below the diagonal in Figure 2C. Data points above the diagonal may reflect novel complexes in which CORUM orthocomplexes and/or subcomplexes interact with unknown proteins. These results are consistent with previous observations (6, 18, 20, 22) and indicate that CORUM subunits detected in CFMS experiments rarely agree with the predicted mass of the fully assembled state.

Gold standard predictions

To identify reliable rice orthocomplexes, we extracted reproducible protein elution profiles from the reference rice SEC datasets (16). There were 3,426 proteins present in both of the two SEC replicates, and 197 had multiple peaks that arose when the protein existed in multiple multimerization states. We deconvolved the multiple peaks to generate 350 reproducible peaks, and a total of 2,618 protein subunits with $R_{app} > 1$ were used as the rice SEC reference profiles. We further curated rice orthocomplexes with a subunit coverage greater than 2/3 in the rice SEC profiles. One hundred and three rice orthocomplexes were selected by integration of rice orthocomplexes into the rice SEC data for further analyses. The small-world analysis was performed across the 103 CORUM orthocomplexes one by one to predict the composition of

CORUM subcomplexes. The results of the small-world analysis, including essential details such as cluster composition, singletons, statistical significance, and size evaluation, were reported in Supplemental Table S3 and illustrated in Supplemental Figures S2 and S3. More specifically, in Supplemental Figure S2, we plotted M_{app} versus M_{calc} of the identified subcomplexes after the small-world analysis, demonstrating the existence of partially-assembled complexes in the rice cell extract. Additionally, SEC profile plots illustrating co-elution patterns of the small-world analysis results were generated in Supplemental Figure S3.

During the process of small-world analysis, subunits within each CORUM orthocomplex were clustered based on their profiles and distances. The outcome of this analysis included 162 subcomplexes/complexes (Figure 3A; Supplemental Figures S2 and S3; Supplemental Table S3), among which 112 had *p*-values less than 5%. After removing redundancies from the list of 112, and discarding potential monomers ($R_{app} \leq 1.6$), we identified 79 unique subcomplexes/complexes with small *p*-values (Supplemental Table S4). We proceeded to compare their M_{app} and M_{calc} values using the criterion defined in Formula (6) to this set of 79 subcomplexes/complexes (Figure 3B; Supplemental Table S4). In comparison to Figure 2C, there were considerably fewer data points below the diagonal, indicating that our algorithm identified more reliable subcomplex formations. Among the set of 79 subcomplexes/complexes, 40 demonstrated substantial agreements between M_{app} and M_{calc}, as shown in Figure 3B, meeting the criterion of Formula (6). These 40 subcomplexes or complexes were stable subcomplexes/complexes supported by both statistically significant *p*-values and consistent apparent masses (Figure 3B and 3C). They serve as gold standards to evaluate CFMS predictions. Additionally, within these 40 gold standards, our algorithm identified 8 fully assembled CORUM orthocomplexes. These fully assembled complexes met the criteria of

containing all CORUM subunits, being statistically significant, and having similar M_{app} and M_{calc} values according to Formula (6). A summary list of these 40 gold standards is provided in Table 1, with additional details in Supplemental Table S5.

Confirmation of monomer identification by the algorithm

Within the set of 103 CORUM orthocomplexes, there existed a list of 50 proteins with small R_{app} values ($R_{app} \leq 1.6$), likely being monomers or multimers with a restricted type of binding partner. This list, derived directly from the original input data, served as prior information to evaluate and validate the accuracy of our algorithm in discerning monomeric proteins within orthocomplexes. Our algorithm identified many of these putative monomers. Specifically, our algorithm correctly predicted 14 putative monomers as singletons, with significantly large distance from the rest of the proteins within their respective CORUM orthocomplexes (Supplemental Table S6A). Furthermore, we examined the scenarios of ortho-paralog subcomplexes, where all rice subunits were mapped to a single human ortholog. Our algorithm accurately identified those ortho-paralog subcomplexes, resulting in the discovery of 9 additional putative monomers (Supplemental Table S6B). The remaining 27 putative monomers were clustered into different subcomplexes by our algorithm, but these subcomplexes had large *p*-values, indicating they were not expected to form discrete complexes.

Structural validation of RNA polymerase II subcomplexes

Many novel subcomplexes were identified through our two-stage clustering approach in this study. The dynamic assembly of RNA polymerase II complex (POL II) with general transcription factors into transcription preinitiation complexes is central to transcriptional control

(40). The rice reference CFMS dataset included reproducible protein elution profiles for 12 subunits of the POL II complex (Figure 4A). RPB3, RPB9, and RBP11a had two reproducible, resolvable peaks. Our clustering assigned the POL II subunits into four different subcomplexes based on their elution profiles and reproducible peaks (Figure 4B). While subunits assigned in the subcomplexes 1, 3, and 4 showed similar profiles in both replicates (Supplemental Figure S3), TFIIB and RPB9 in the subcomplex 2 had a peak at around fractions 18-19. Thus, each subcomplex prediction was evaluated using a statistical bootstrap *p*-value calculation (Supplemental Table S3). As shown in the profiles, all the subunits within the subcomplexes 1, 3, and 4 had predicted *p*-values less than 0.01. Our analysis predicts that the low and high mass peaks of RBP3 and RBP11 correspond to a heterodimer and subcomplex 3, respectively. The prediction for the subcomplex 2 was insignificant (*p*-value = 0.56), and both subunits were predicted to be monomeric based on $R_{app} \leq 1.6$.

We also compared the M_{calc} values of predicted subcomplexes and M_{app} values of subunits of the subcomplexes to evaluate the predictions (Figure 4B). In addition to their significant *p*-values, M_{app} values of predicted subcomplex subunits were plotted nearby, supporting the presence of those three significant subcomplexes in the reference datasets. The slight skewness toward elevated M_{app} might be due to undetected or unknown proteins in the complex with non-spherical shapes. TFIIB in both replicates had similar peak locations (subcomplex 2 Fractions 12-15) to RPB3 and RBP11 peaks in the subcomplex 1 (Figure 4A and Supplemental Figure S3), indicating a potential association of TFIIB to the subcomplex 1.

To test for potential direct interactions, pairwise interactions among the subunits in subcomplexes 1 and 2 were analyzed using AlphaFold Multimer (32). The mean of the confidence scores from the top 5 predicted models were calculated for the 10 possible

combinations (Figure 4C). The heteromerization between RPB3 and RBP11 possessed the highest model confidence (ranking-confidence score: 0.87 ± 0.01), supporting the predicted subcomplex 1. All other heteromeric and homomeric models for TFIIB were also predicted with very low ranking-confidence scores (0.16 ± 0.01 ~ 0.31 ± 0.01), indicating TFIIB may not be retained as a stable complex with other RNA pol II subunits in the cell extracts. The predicted three subcomplexes were also structurally validated (Figure 4D). The Cryo-EM solved structure of PIC (40) was downloaded from the Protein Data Bank (PDB: 6O9L). All the subunits in the predicted three subcomplexes are mapped onto the holoenzyme structure in a spatially plausible configuration (Figure 4D(1)). Subunits of the subcomplex 4 were assigned to the TFIIH complex (Figure 4D(2)), while those of the subcomplex 3 were assigned to the POL II complex (Figure 4D(3)). Each of them was also supported by the solved structures (PDB: 6DRD & 7NVW). The second peaks of RPB3 and RBP11 assembled into the subcomplex 1 with size consensus M_{calc} and M_{app}. RPB3 and RPB11 heteromerization has been shown in Arabidopsis (41), and the subcomplex is at the core of POLII assembly (42). Our data are consistent with a model in which the RPB3 and RPB11 heteromerization occurs prior to the association with RPB10 and RPB12 to form the RPB3 subcomplex (43, 44). In summary, this example provides structural validation supporting our subcomplex predictions, pointing to the existence of discrete RNA Pol II subcomplexes in the cell. It seems common that a full holocomplex assembly is a regulated event, and CFMS data can provide clues about the path through which this occurs. Interestingly, these abundant sub-complexes may not reside in the nucleus. The cytosolic fraction analyzed in this study is not enriched in abundant nuclear proteins like histones (6, 22), and cytosolic assembly of core RNA POL II has been demonstrated in a human cell line (43-45). Therefore, these subcomplexes could reflect entities with distinct functions in the cytosol and/or protein

complexes that cycle between nuclear and non-nuclear localizations.

Importantly, it is worth noting that our methods and results are not specific to plants. Partial assembly of CORUM complexes likely occurs in many organisms, and our methods can be directly applied to detect it in other species.

Use of gold standards to evaluate a global protein complex prediction

The gold standards were used to evaluate the global prediction results in the rice reference protein complex datasets (16). In the published protein complex prediction, the resulting dendrogram at a specific cluster number determines the compositions of predicted complexes. The specific cluster number is selected based on the resolution of the data and maximized to decrease false positives (23). Of our 40 gold standards, 34 were present in the reference SEC and IEX data used for the protein complex predictions, and they consisted of multiple subunits. Consequently, we employed them to assess the overall reliability of 1,000 predicted protein complexes derived from the rice dataset (16). The assessment was performed using the intactness and purity metrics, as defined in Formula (7). These metrics play a vital role in assessing the quality of complex predictions. In a reliable prediction, gold standard complexes should exhibit high intactness and purity. On the other hand, these two metrics are affected by the size of the protein complexes (see Materials and Methods). Figure 5 displays the intactness and purity values, comparing our gold standard subcomplexes and their fully assembled CORUM counterparts. In the assessment of specific global protein complex predictions, the intactness values of our gold standards exceeded those of the CORUM full orthocomplexes (Figure 5A), while the purity values were comparable between the two (Figure 5B). As our gold standards tend to be smaller than the CORUM full complexes, and since the size of reference

complexes impacts the intactness and purity values in opposite directions, it is essential to consider these two metrics together. The higher intactness values and similar purity values collectively indicate that the global predicted protein complexes align better with our gold standards than with the CORUM full complexes.

Our gold standards, which are essentially predicted knowns and validated by statistical significance and matched size masses from SEC experiments (M_{app}) and the calculation (M_{calc}), have shown robust performance in evaluating protein complex predictions. While higher intactness or purity values are typically deemed optimal with the use of confirmed known complexes as references, our gold standards affirm their utility when computationally validated. These gold standards provide a reliable benchmark for assessing protein complex predictions.

Conclusion

In the living cell, fully assembled CORUM complexes reflect an active state based on a wide variety of genetic and biochemical data. However, this does not mean that the fully assembled complex is the most abundant state of the subunits in cells, nor does it exclude the possibility that subcomplexes have functions that are independent of the fully assembled complex. In plant cell extracts, CORUM orthocomplexes are rarely fully assembled (6, 22) (Figure 2C), and to our knowledge, this has not been addressed directly in non-plant systems. The method described here provides novel statistical approaches to identify a refined set of gold standards. Broad adoption of these methods will enable more accurate evaluation of protein complex predictions and the more reliable use of machine learning methods to improve CFMS-based predictions of protein complex composition.

Funding

This work was supported by the National Science Foundation (NSF) Plant Genome Research Project 1951819 to D.B.S.

References

1. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 426, 570-574

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

3. Havugimana, Pierre C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, Andrei L., Li, Z., Wang, Peggy I., Boutz, Daniel R., Fong, V., Phanse, S., Babu, M., Craig, Stephanie A., Hu, P., Wan, C., Vlasblom, J., Dar, V.-u.-N., Bezginov, A., Clark, Gregory W., Wu, Gabriel C., Wodak, Shoshana J., Tillier, Elisabeth R. M., Paccanaro, A., Marcotte, Edward M., and Emili, A. (2012) A census of human soluble protein complexes. *Cell* 150, 1068-1081

4. Aryal, U. K., Xiong, Y., McBride, Z., Kihara, D., Xie, J., Hall, M. C., and Szymanski, D. B. (2014) A proteomic strategy for global analysis of plant protein complexes. *The Plant Cell* 26, 3867

5. Aryal, U. K., McBride, Z., Chen, D., Xie, J., and Szymanski, D. B. (2017) Analysis of protein complexes in *Arabidopsis* leaves using size exclusion chromatography and label-free protein correlation profiling. *Journal of Proteomics* 166, 8-18

6. McBride, Z., Chen, D., Reick, C., Xie, J., and Szymanski, D. B. (2017) Global analysis of membrane-associated protein oligomerization using protein correlation profiling. *Molecular & Cellular Proteomics*, mcp.000276.002017

7. Gilbert, M., and Schulze, W. X. (2019) Global identification of protein complexes within the membrane proteome of Arabidopsis roots using a SEC-MS approach. *Journal of Proteome Research* 18, 107-119

8. Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., Samaras, P., Richter, S., Shikata, H., Messerer, M., Lang, D., Altmann, S., Cyprys, P., Zolg, D. P., Mathieson, T., Bantscheff, M., Hazarika, R. R., Schmidt, T., Dawid, C., Dunkel, A., Hofmann, T., Sprunck, S., Falter-Braun, P.,

Johannes, F., Mayer, K. F. X., Jürgens, G., Wilhelm, M., Baumbach, J., Grill, E., Schneitz, K., Schwechheimer, C., and Kuster, B. (2020) Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* 579, 409-414

9. McWhite, C. D., Papoulas, O., Drew, K., Cox, R. M., June, V., Dong, O. X., Kwon, T., Wan, C., Salmi, 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.e414

10. Olinares, P. D. B., Ponnala, L., and van Wijk, K. J. (2010) Megadalton Complexes in the Chloroplast Stroma of Arabidopsis thaliana Characterized by Size Exclusion Chromatography, Mass Spectrometry, and Hierarchical Clustering*. *Molecular & Cellular Proteomics* 9, 1594-1615

11. Senkler, J., Senkler, M., Eubel, H., Hildebrandt, T., Lengwenus, C., Schertl, P., Schwarzländer, M., Wagner, S., Wittig, I., and Braun, H.-P. (2017) The mitochondrial complexome of Arabidopsis thaliana. *The Plant Journal* 89, 1079-1092

12. Klodmann, J., Sunderhaus, S., Nimtz, M., JÄnsch, L., and Braun, H.-P. (2010) Internal Architecture of Mitochondrial Complex I from Arabidopsis thaliana. *The Plant Cell* 22, 797-810

13. Gorka, M., Swart, C., Siemiatkowska, B., Martínez-Jaime, S., Skirycz, A., Streb, S., and Graf, A. (2019) Protein Complex Identification and quantitative complexome by CN-PAGE. *Scientific Reports* 9, 11523

14. Veyel, D., Kierszniowska, S., Kosmacz, M., Sokolowska, E. M., Michaelis, A., Luzarowski, M., Szlachetko, J., Willmitzer, L., and Skirycz, A. (2017) System-wide detection of protein-small molecule complexes suggests extensive metabolite regulation in plants. *Scientific Reports* 7, 42387

15. Veyel, D., Sokolowska, E. M., Moreno, J. C., Kierszniowska, S., Cichon, J., Wojciechowska, I., Luzarowski, M., Kosmacz, M., Szlachetko, J., Gorka, M., Méret, M., Graf, A., Meyer, E. H., Willmitzer, L., and Skirycz, A. (2018) PROMIS, global analysis of PROtein–metabolite interactions using size separation in Arabidopsis thaliana. *Journal of Biological Chemistry* 293, 12440-12453

16. Lee, Y., Okita, T. W., and Szymanski, D. B. (2021) A co-fractionation mass spectrometry-based prediction of protein complex assemblies in the developing rice aleurone-subaleurone. *The Plant Cell* 33,

2965-2980

17. Salas, D., Stacey, G. R., Akinlaja, M., and Foster, L. J. (2020) Next-generation Interactomics: Considerations for the use of co-elution to measure protein interaction networks. *Molecular & Cellular Proteomics* 19, 1

18. Pang, C. N. I., Ballouz, S., Weissberger, D., Thibaut, L. M., Hamey, J. J., Gillis, J., Wilkins, M. R., and Hart-Smith, G. (2020) Analytical Guidelines for co-fractionation Mass Spectrometry Obtained through Global Profiling of Gold Standard Saccharomyces cerevisiae Protein Complexes. *Molecular & Cellular Proteomics* 19, 1876-1895

19. Heide, H., Bleier, L., Steger, M., Ackermann, J., Dröse, S., Schwamb, B., Zörnig, M., Reichert, Andreas S., Koch, I., Wittig, I., and Brandt, U. (2012) Complexome Profiling Identifies TMEM126B as a Component of the Mitochondrial Complex I Assembly Complex. *Cell Metabolism* 16, 538-549

20. Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A., Chessman, K., Pal, S., Cromar, 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., R. Tillier, E., Wallingford, J. B., Parkinson, J., Marcotte, E. M., and Emili, A. (2015) Panorama of ancient metazoan macromolecular complexes. *Nature* 525, 339-344

21. Heusel, M., Bludau, I., Rosenberger, G., Hafen, R., Frank, M., Banaei-Esfahani, A., van Drogen, A., Collins, B. C., Gstaiger, M., and Aebersold, R. (2019) Complex-centric proteome profiling by SEC-SWATH-MS. *Molecular Systems Biology* 15, e8438

22. Lee, Y., and Szymanski, D. B. (2021) Multimerization variants as potential drivers of neofunctionalization. *Science Advances* 7, eabf0984

23. McBride, Z., Chen, D., Lee, Y., Aryal, U. K., Xie, J., and Szymanski, D. B. (2019) A label-free mass spectrometry method to predict endogenous protein complex composition. *Molecular & Cellular Proteomics* 18, 1588

24. Heusel, M., Frank, M., Köhler, M., Amon, S., Frommelt, F., Rosenberger, G., Bludau, I., Aulakh, S., Linder, M. I., Liu, Y., Collins, B. C., Gstaiger, M., Kutay, U., and Aebersold, R. (2020) A Global

Screen for Assembly State Changes of the Mitotic Proteome by SEC-SWATH-MS. *Cell Systems* 10, 133-155.e136

25. Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2018) CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research* 47, D559-D563

26. Fossati, A., Li, C., Uliana, F., Wendt, F., Frommelt, F., Sykacek, P., Heusel, M., Hallal, M., Bludau, I., Capraz, T., Xue, P., Song, J., Wollscheid, B., Purcell, A. W., Gstaiger, M., and Aebersold, R. (2021) PCprophet: a framework for protein complex prediction and differential analysis using proteomic data. *Nature Methods* 18, 520-527

27. Sonnhammer, E. L. L., and Östlund, G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research* 43, D234-D239

28. Clark, J. W., and Donoghue, P. C. J. (2018) Whole-genome duplication and plant macroevolution. *Trends in Plant Science* 23, 933-945

29. de Gelder, R., Wehrens, R., and Hageman, J. A. (2001) A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *Journal of Computational Chemistry* 22, 273-289

30. Kohonen, T. (1990) The self-organizing map. *Proceedings of the IEEE* 78, 1464-1480

31. Frey, B. J., and Dueck, D. (2007) Clustering by Passing Messages Between Data Points. *Science* 315, 972-976

32. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2022) Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034

33. Cianfrocco, M. A., Wong-Barnum, M., Youn, C., Wagner, R., and Leschziner, A. (2017) COSMIC2: A Science Gateway for Cryo-Electron Microscopy Structure Determination. *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and*

Impact, p. Article 22, Association for Computing Machinery, New Orleans, LA, USA

34. Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., and Buell, C. R. (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Research* 35, D883-D887

35. Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2011) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40, D1178-D1186

36. R Core Team (2018) R: A language and environment for statistical computing. 3.5.1 Ed., R Foundation for Statistical Computing, Vienna, Austria

37. RStudio Team (2018) RStudio: Integrated development environment for R. 1.1.463 Ed., RStudio, Inc., Boston, MA

38. Wehrens, R., and Kruisselbrink, J. (2018) Flexible Self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software* 87, 1 - 18

39. Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics* 27, 2463-2464

40. Yan, C., Dodd, T., He, Y., Tainer, J. A., Tsutakawa, S. E., and Ivanov, I. (2019) Transcription preinitiation complex structure and dynamics provide insight into genetic diseases. *Nature Structural & Molecular Biology* 26, 397-406

41. Ulmasov, T., Larkin, R. M., and Guilfoyle, T. J. (1996) Association between 36- and 13.6-kDa α-Like Subunits of Arabidopsis thaliana RNA Polymerase II (∗). *Journal of Biological Chemistry* 271, 5085-5094

42. Acker, J., de Graaff, M., Cheynel, I., Khazak, V., Kedinger, C., and Vigneron, M. (1997) Interactions between the Human RNA Polymerase II Subunits*. *Journal of Biological Chemistry* 272, 16815-16821

43. Garrido-Godino, A. I., Gutiérrez-Santiago, F., and Navarro, F. (2021) Biogenesis of RNA Polymerases in Yeast. *Frontiers in Molecular Biosciences* 8

44. Wild, T., and Cramer, P. (2012) Biogenesis of multisubunit RNA polymerases. *Trends in Biochemical Sciences* 37, 99-105

45. Boulon, S., Pradet-Balade, B., Verheggen, C., Molle, D., Boireau, S., Georgieva, M., Azzag, K., Robert, M.-C., Ahmad, Y., Neel, H., Lamond, A. I., and Bertrand, E. (2010) HSP90 and Its R2TP/Prefoldin-like Cochaperone Are Involved in the Cytoplasmic Assembly of RNA Polymerase II. *Molecular Cell* 39, 912-924

Table 1. Predicted subcomplexes that could be used as gold standards to evaluate CFMS based protein complex predictions.

Figure legends

Figure 1. Identification of bona-fide gold standards to evaluate prediction accuracies in co-fractionation mass spectrometry-based protein complex discovery in plant species. Subcomplex predictions in SEC datasets are evaluated via a statistical bootstrap *p*-value calculation. Among experimentally detected subunits (green) in a CORUM orthocomplex, subunits with similar SEC profiles are clustered in a subcomplex (dotted line). Profile similarity scores are calculated between all possible pairs of subunits in the subcomplex and then are averaged to get the mean dissimilarity of the subcomplex. At the same time, the same number of proteins observed in the orthocomplex are sampled from randomly generated plant orthocomplex (yellow). The random mean is calculated as mean dissimilarity for pairs of proteins in the random subcomplex. The *p*-value for each subcomplex is calculated as the fraction of times the observed mean is larger than the random mean.

Figure 2. Assumed CORUM gold standard complexes do not exist in a fully assembled state in plant species. A, Genomic level subunit coverages of CORUM complexes to rice orthocomplexes. The coverages are defined as the ratio of the number of subunits in a rice orthocomplex to the number of subunits in its orthologous human CORUM complexes. The genome coverages of 1964 rice orthocomplexes were calculated and plotted at different subunit coverages. **B,** Conserved predicted masses of human complexes and rice orthocomplexes. M_{calc} of 436 rice orthocomplexes with 100% subunit coverage to CORUM complexes were plotted. **C,**

CORUM complex subunits rarely exist in a fully assembled complex. A scatter plot presents protein complex conservation in size between CORUM complexes and rice orthocomplexes. M_{app} values of subunits of the 258 rice orthocomplexes were obtained from the reference rice CFMS datasets. **D,** Conserved masses of human CORUM subunits and rice orthocomplex subunits. A scatter plot shows conserved M_{mono} values between human and rice subunits that assemble into the known complexes. The rice orthocomplexes with 100% subunit coverage to CORUM complexes were plotted.

Figure 3. Useful gold standards near the diagonal. A, A process flow to identify gold standards from the small world analysis result. **B,** Gold standard subcomplexes/complexes with matched M_{calc} and $M_{app\text{-}avg}$ are rendered in pink. Asterisk (*) indicates fully assembled CORUM orthocomplexes. Numbers in parentheses point out predicted subcomplexes present in the corresponding panel in Figure 3C. **C,** M_{calc} values of predicted subcomplexes and M_{app} values of subunits of the subcomplexes. Gold standard subcomplexes (*p*-values = 0.0) were highlighted using bold text font.

Figure 4 Validation of subcomplexes predicted in CORUM RNA polymerase II complex. A, Protein elution profiles of subunits in each predicted subcomplex. Profiles in another replicate can be found in the Supplemental Figure S3. **B,** M_{calc} values of predicted subcomplexes and M_{app} values of subunits of the subcomplexes. Circles indicate multimers, while triangles mean monomer (R_{app} < 1.6). **C,** Dimerization prediction between subcomplex subunits by AlphaFold Multimer (32). AlphaFold Multimer was run on COSMIC[2] to predict top5 models (33). Each value in the table is the mean of ranking confidence score (ipTM + pTM)□±□standard

deviation. **D,** Structural validation of predicted subcomplexes. Predicted subcomplexes are searched in RCSB Protein Data Bank (https://www.rcsb.org/). The structure of the fully assembled CORUM RNA polymerase II complex is available (PDB: 6O9L). (1) Undetected subunits in the CFMS dataset are colored gray. (2) – (4) Subcomplexes were predicted in B. M_{calc} and M_{app} of predicted rice subcomplexes are summarized right next to the subcomplex structures.

Figure 5. The validated subcomplexes and CORUM complexes were used as standards to evaluate a global prediction of protein complex composition in rice. The global predictions of rice complexes (16) were re-evaluated with respect to the intactness (A) and purity (B) of assumed fully assembled CORUM orthocomplexes and the validated gold standards defined in this study.

Supplemental Figures

Supplemental Figure S1. Subunit coverages of rice orthocomplexes.

Supplemental Figure S2. Subcomplex identification by small world analysis (M_{app} vs M_{calc} plots).

Supplemental Figure S3. Protein SEC profiles illustrating small-world analysis results.

Supplemental Tables

Supplemental Table S1. Human to rice ortholog mapping.

Supplemental Table S2. CORUM human complexes and rice orthocomplexes.

Supplemental Table S3. The list of 103 rice orthocomplexes and their subcomplex prediction results.

Supplemental Table S4. Significant subcomplexes with small *p*-values.

Supplemental Table S5. Predicted subcomplexes that could be used as gold standards to evaluate CFMS based protein complex predictions.

Supplemental Table S6. Predicted monomers.

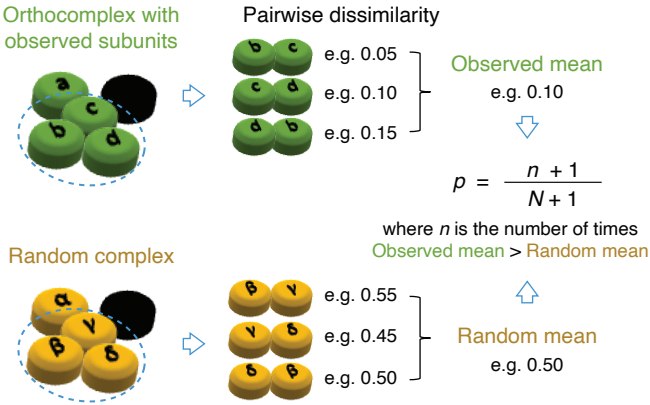
Creating CORUM-to-Rice orthocomplexes

Co-fraction mass spectrometry (CFMS)

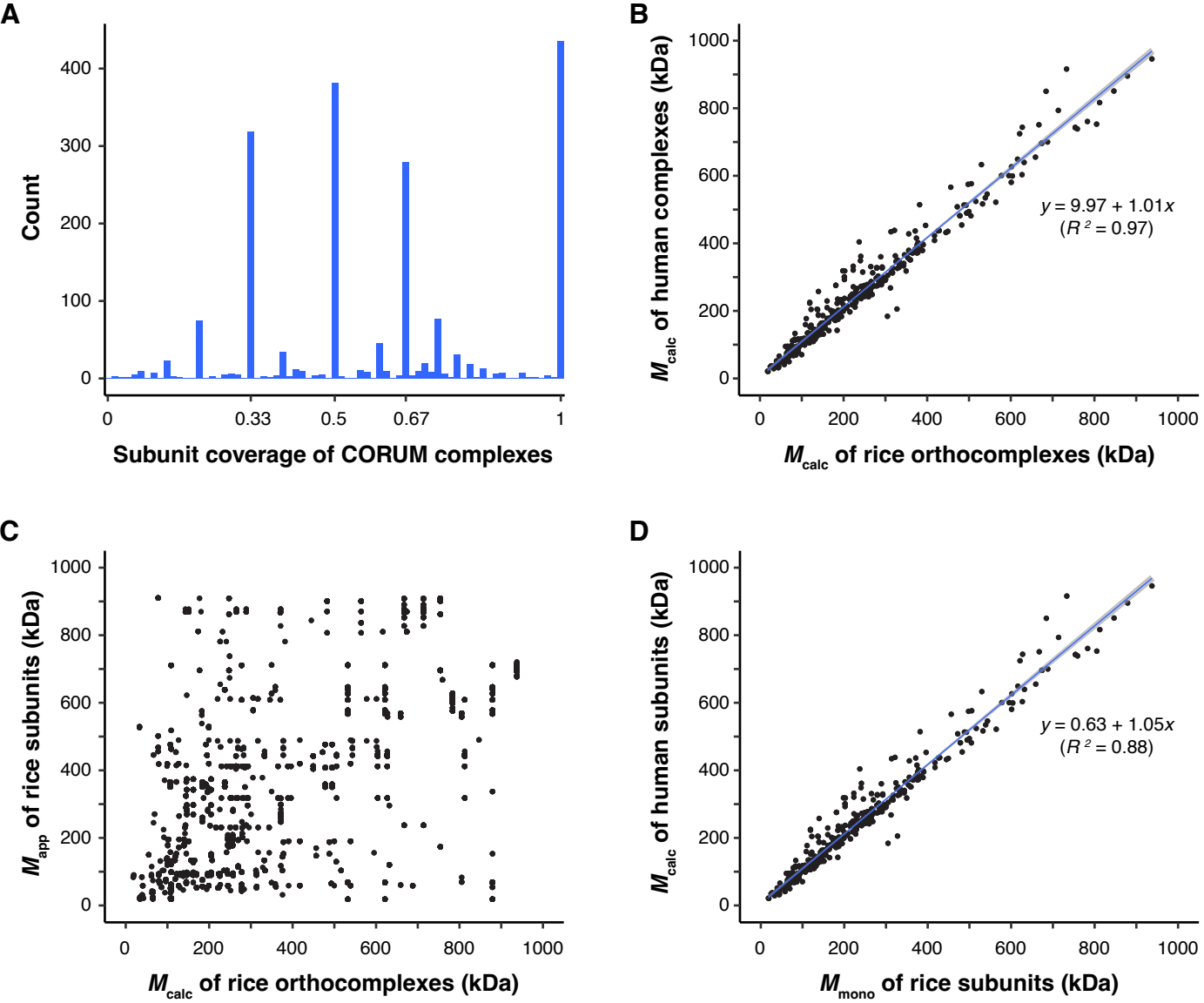
- LC/MSMS
- Reproducible peak filtering
- Gaussian peak detection

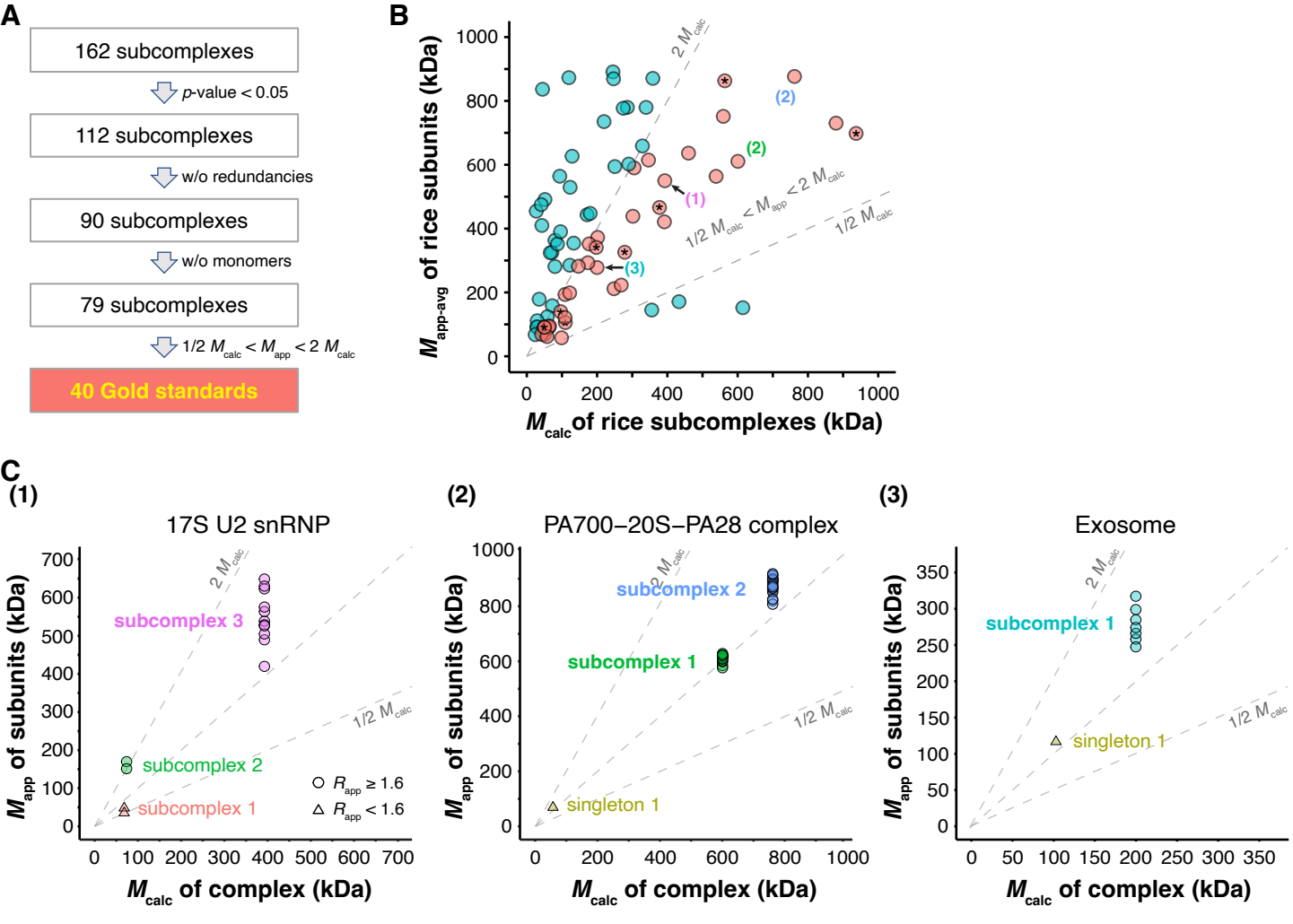
Clustering by Self Organizing Map (SOM)

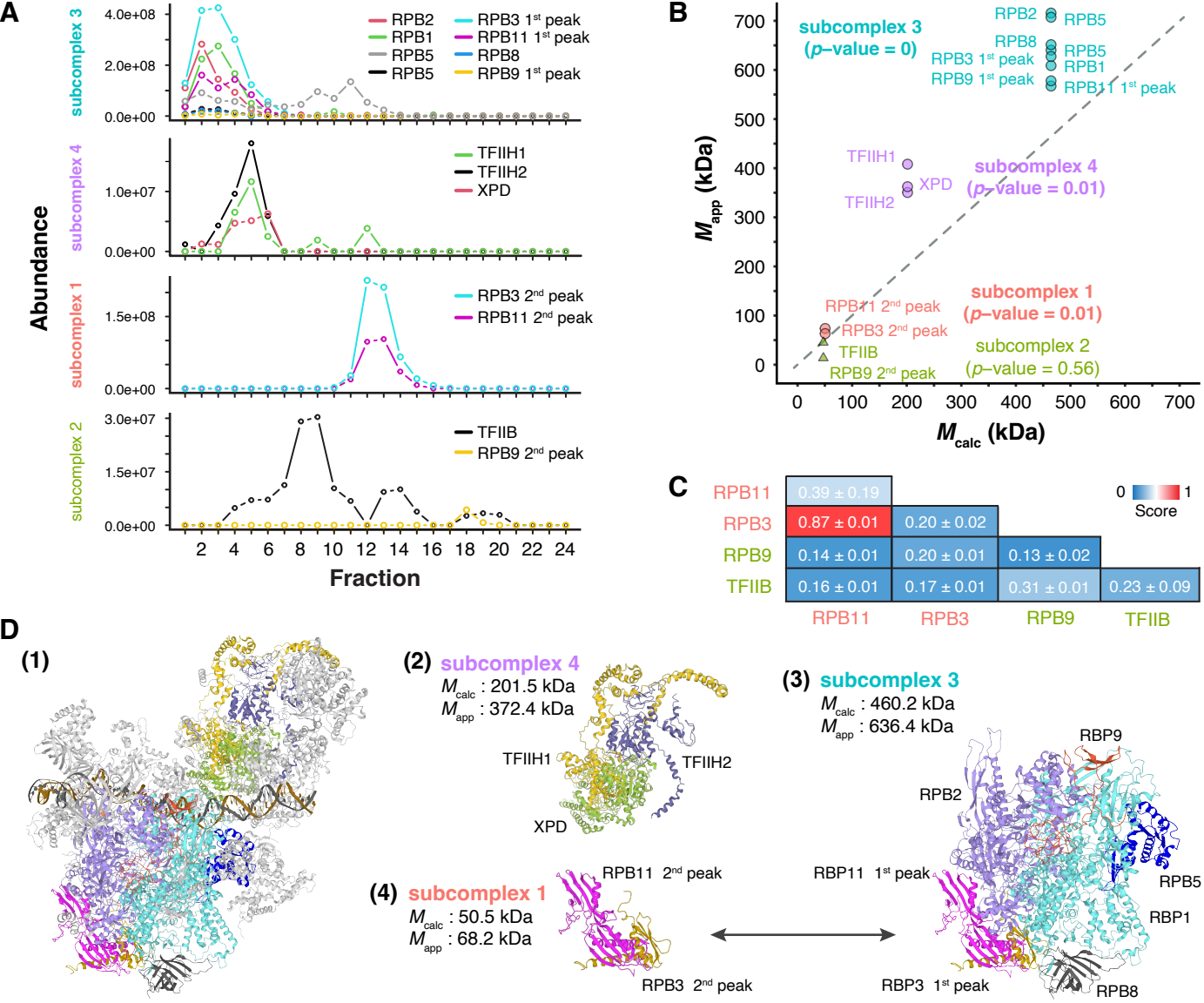
Evaluation of subcomplex predictions

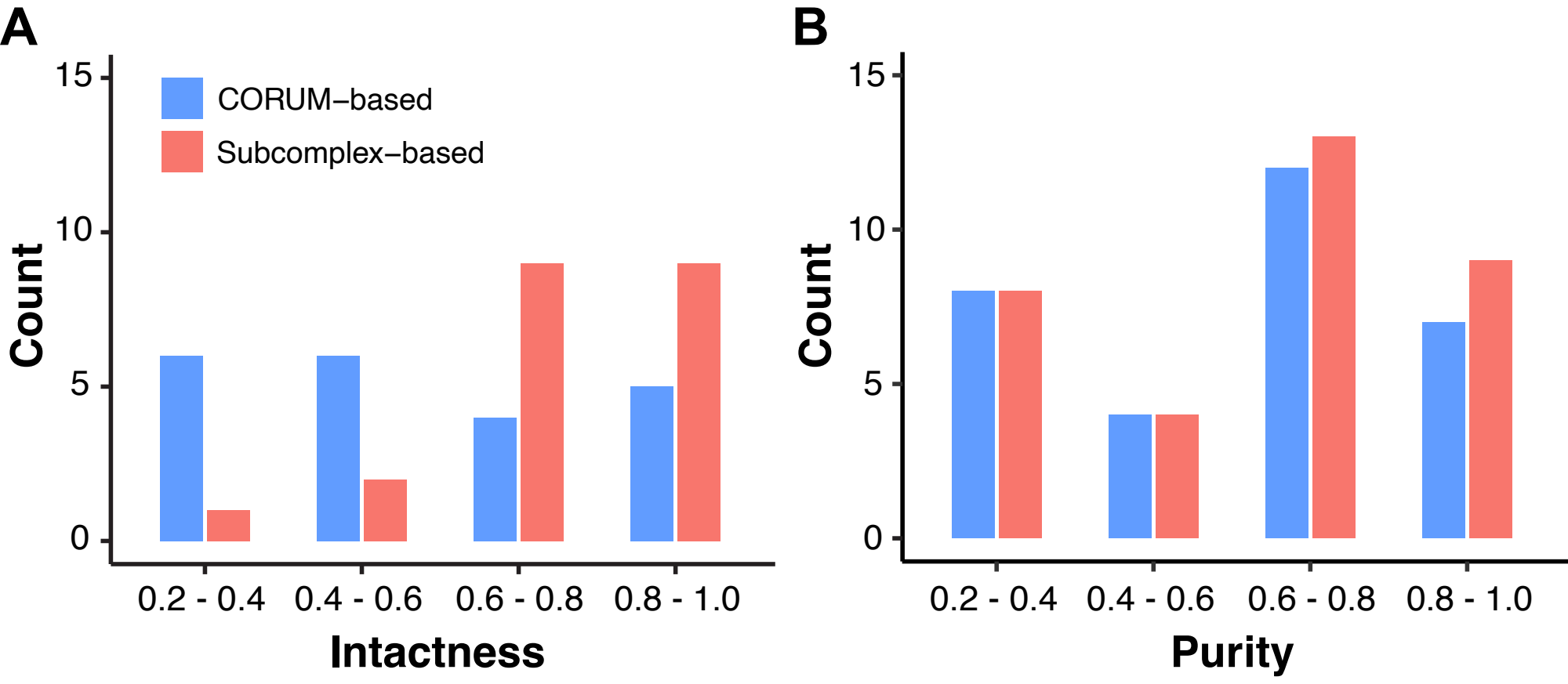


Golden Standards for CFMS


A
Count
0
100
200
300
400
0
0.33
0.5
0.67
1
Subunit coverage of CORUM complexes
B
M_{calc} of human complexes (kDa)
M_{calc} of rice orthocomplexes (kDa)
$y = 9.97 + 1.01x$
($R^2 = 0.97$)
C
M_{app} of rice subunits (kDa)
M_{calc} of rice orthocomplexes (kDa)
D
M_{calc} of human subunits (kDa)
M_{mono} of rice subunits (kDa)
$y = 0.63 + 1.05x$
($R^2 = 0.88$)


A
162 subcomplexes
p-value < 0.05
112 subcomplexes
w/o redundancies
90 subcomplexes
w/o monomers
79 subcomplexes
1/2 M_calc < M_app < 2 M_calc
40 Gold standards
B
M_app-avg of rice subunits (kDa)
M_calc of rice subcomplexes (kDa)
2 M_calc
1/2 M_calc < M_app < 2 M_calc
1/2 M_calc
(1)
(2)
(3)
C
(1)
17S U2 snRNP
subcomplex 3
subcomplex 2
subcomplex 1
R_app ≥ 1.6
R_app < 1.6
M_app of subunits (kDa)
M_calc of complex (kDa)
(2)
PA700–20S–PA28 complex
subcomplex 2
subcomplex 1
singleton 1
(3)
Exosome
subcomplex 1
singleton 1


A
Abundance
subcomplex 3
4.0e+08
2.0e+08
0.0e+00
RPB2
RPB1
RPB5
RPB5
RPB3 1st peak
RPB11 1st peak
RPB8
RPB9 1st peak
subcomplex 4
1.0e+07
0.0e+00
TFIIH1
TFIIH2
XPD
subcomplex 1
1.5e+08
0.0e+00
RPB3 2nd peak
RPB11 2nd peak
subcomplex 2
3.0e+07
1.5e+07
0.0e+00
TFIIB
RPB9 2nd peak
2 4 6 8 10 12 14 16 18 20 22 24
Fraction
B
subcomplex 3
(p–value = 0)
RPB2
RPB5
RPB8
RPB5
RPB3 1st peak
RPB1
RPB9 1st peak
RPB11 1st peak
TFIIH1
XPD
TFIIH2
subcomplex 4
(p–value = 0.01)
RPB11 2nd peak
RPB3 2nd peak
TFIIB
RPB9 2nd peak
subcomplex 1
(p–value = 0.01)
subcomplex 2
(p–value = 0.56)
Mapp (kDa)
Mcalc (kDa)
0 100 200 300 400 500 600 700
C
RPB11
RPB3
RPB9
TFIIB
0.39 ± 0.19
0.87 ± 0.01
0.20 ± 0.02
0.14 ± 0.01
0.20 ± 0.01
0.13 ± 0.02
0.16 ± 0.01
0.17 ± 0.01
0.31 ± 0.01
0.23 ± 0.09
0 1
Score
D
(1)
(2) subcomplex 4
Mcalc : 201.5 kDa
Mapp : 372.4 kDa
TFIIH1
TFIIH2
XPD
(3) subcomplex 3
Mcalc : 460.2 kDa
Mapp : 636.4 kDa
RBP9
RPB2
RBP11 1st peak
RPB5
RBP1
RBP3 1st peak
RPB8
(4) subcomplex 1
Mcalc : 50.5 kDa
Mapp : 68.2 kDa
RPB11 2nd peak
RPB3 2nd peak


A
B
CORUM–based
Subcomplex–based
Count
15
10
5
0
0.2 - 0.4
0.4 - 0.6
0.6 - 0.8
0.8 - 1.0
Intactness
Purity