Supplementary Information

Generating the Interaction network from the raw pull-down mass spec data

Networks built in this study from the experimentally observed interactions:

Network	#genes/nodes	# interactions/edges
Extended Network	3672	14317
Core Network	2708	7123
Merged Network	2186	5496
Intersection Network	1210	2357

The extended network includes all interactions with scores ≥0.101
The core network is derived from the same network as above but includes only those interactions with scores ≥0.273. In all 4 networks interactions with ribosomal prey proteins are not included.

The Intersection and Merged networks represent more naïve ways of combining data and scores from the Maldi and LC/MS experiments. They include respectively, all interactions identified in both experiments, and all the interactions identified in both experiments plus those reciprocally (bait-pray/pray-bait) and those repeatedly (2x bait-pray) identified in either method. The score threshold for including interactions was 70% for LC/MS/MS and 1.0 for the Maldi Z-score. The integrated interaction scores were computed as follows: for LC/MS/MS the percentage scores were converted to Z-scores and then rescaled such that a value of 0.5 corresponds to 70%. For Maldi, the Z-scores were rescaled so that a value 0.5 corresponds to original Z-score of 1. This was done in order to make the scores of both methods comparable to one another. The final score of an interaction was computed as the average of the rescaled Z-scores from multiple experiments.

Visualization and analysis of complexes using Cytoscape

Cytoscape is a public domain software environment in Java for the analysis and visualization of biomolecular interaction networks[1]. The Cytoscape software core provides basic functionality to layout and query networks; to visually integrate a network with various types of data such as expression levels, functional annotations, phenotypes; and to link displayed items (nodes and edges) to external databases with additional information. The Core is extensible through a straightforward plug-in architecture, allowing rapid development of additional computational analyses and features. GenePro is a 'plug-in' that provides several integrative and interactive visualization and analysis capabilities for networks of interacting proteins and genes to be described elsewhere (Orsi et al., unpublished).

The highly connected modules identified by the MCL clustering procedure (see below and main text for details), which we take here to represent multi-protein complexes characterized in this study, are displayed and analyzed using the GenePro Cytoscape plug-in. Figures 3d and S3 display the identified complexes in the context of the global 'core network' from which they are derived. Each complex is represented as a node, and 2 nodes are linked by an edge whenever proteins in one node form at least 2 interactions with proteins in the other node, with the thickness of the edge being proportional to the number of observed interactions between the connected nodes. This representation positions the identified protein complexes within the interaction network and shows that, although the complexes represent highly connected modules where the proteins form many interactions with one another, some interactions are also formed between proteins in different modules. This indicates that the assignment of these proteins to specific modules can be arbitrary, and these proteins could just as well be part of two or more complexes. The number of such shared proteins in the set of complexes identified here is, however, limited. Positioning the Mouse over an edge linking two complexes displays the number of individual protein-protein interactions between these complexes. A Mouse-over any node/complex provides information about various properties of the proteins in the complex.

Some of these properties are summarized graphically (Fig. S3). Each complex is displayed as a pie chart, whose size reflects the size of the complex. The size and colour of each section of the pie represent the fraction of the proteins in each complex that map into a given complex from the hand-curated complexes in the MIPS database[2]. A Mouse-over a given pie section lists the names of all the proteins in the complex identified here that map into the same MIPS complex, and a Mouse click over the same pie section will highlight proteins in other nodes anywhere in the network, which also belong to the same complex. This enables the user to check if proteins from one MIPS complex map into one or more or our complexes. A similar display can be generated highlighting instead the cellular localization of proteins in our complexes or GO functional annotations.

Our GenePro plug-in enables detailed analysis of proteins and interactions within individual complexes. A left Mouse double-click on a complex node displays a new network, which represents all the proteins in the complex as nodes of one colour (red in Fig S3 inset) and all the observed interactions as the edges between these nodes. A Mouse click over an edge displays a small Table which lists the raw reliability scores for each interaction output by the experimental procedures (LC/MS or Maldi, or both) and the reliability measure (or edge weight) derived from these scores that was finally assigned to it. In addition to the proteins within a complex and their interactions, we also display their first degree neighbors in the protein-protein interaction network. These neighbours are defined as any protein outside the complex (most of which are assigned to other complexes) making one or more interactions with a protein inside the complex.

These proteins are displayed as nodes of a different colour from the proteins within the complex (blue in Fig. S3). All displayed intra- and inter-complex edges are coloured according to the value of their weight, so that the user can readily distinguish highly reliable interactions from less reliable ones. A Mouse-over any node displays the gene/protein name, and a Mouse-over any edge displays the edge weight. A Mouse click on a protein/node opens a menu with links to other databases, such as the MIPS, enabling querying for information on that gene.

Deriving complexes from the interaction network

Identifying the multi-protein complexes purified by the experimental procedures involved identifying highly connected modules within the global interaction network. This can be achieved with the help of appropriate clustering procedures[3-5]. Here we use the Markov Cluster (MCL) algorithm, which simulates random walks within graphs using the language of Markov (stochastic) matrices[6] in order to partition a graph into highly connected modules. This procedure handles weighted graphs and works efficiently on large dense graphs, where it displays good convergence and robustness. The excellent performance of MCL relative to other available clustering procedures was demonstrated in a recent study (Brohée et al., unpublished data), which systematically evaluated the ability of several clustering algorithms to identify meaningful modules of densely interacting proteins in a large protein-protein interaction graph.

In the present study the MCL algorithm is applied to the core network and to two other networks derived from the same experimental data but using a more naïve approach, namely the ‘intersection network’ and the ‘merged network’ computed as described above. In each case, we tested several values for the 2 adjustable parameters of this procedure, respectively, the *expansion* and *inflation* operators, settling on the values which provided the best overlap of the computed clusters with the hand-curated complexes from the MIPS database.

The complexes identified by the MCL procedure were then mapped back onto the original protein-protein interaction network from which they were derived. This allowed us to retrieve high confidence interactions between proteins in different complexes. Those interactions were found to occur for a subset of the complexes. Assignment of such connecting proteins to a particular complex may therefore be arbitrary and this limited number of proteins could just as well be part of two or more distinct complexes. The subset of complexes engaged in such high confidence inter-complex interactions can therefore be considered as overlapping. Thus, combining the non-overlapping complexes identified by the MCL algorithm with information on the experimentally derived pairwise interactions can provide equivalent information to that of a clustering procedure that generates overlapping modules.

Quality assessment of derived complexes:

In order to assess the quality of the complexes derived by applying our clustering procedure to the interaction network built from the experimentally determined interactions the following analyses were performed: 1) the correspondence between the complexes derived in this work and the hand-curated complexes from the MIPS database[2] was evaluated, 2) the semantic similarity scores within complexes considering the process taxonomy of GO were computed, 3) we mapped information on cellular localization onto the complexes. In the following, details of these analyses are provided.

All three quality scores were also computed for 1000 randomized networks, having exactly the same connectivity and topology, in order to evaluate statistical significance. This allowed us to derived P-values for the 3 quality scores. In all cases (all the non random datasets analyzed here and for all quality measures), these P-values were extremely low with the highest values equaling 1.7E-87.

1) Evaluating the overlap of computed complexes with those in MIPS

The overlap with the MIPS complexes was evaluated using the measures derived in the study of Brohée et al. (unpublished data). Considering the C_1....C_n complexes/clusters computed in this work, and the M_1.....M_m complexes from the MIPS database, we computed a Confusion Table. Each entry of the Table lists the number of proteins in common between an individual cluster C_j and a MIPS complex M_i .The rows (i) of this Table thus list how the proteins from each of the M_i complexes in MIPS are distributed among the Cj complexes or clusters derived here. Its columns (j) list how the proteins from the Cj complexes are distributed among the MIPS complexes. For each MIPS complex M_i we then compute 2 quantities (Brohée et al., unpublished data):

$$S_i = \max_j(P_{ij}) / \sum_{j=1}^{n} P_{ij} \qquad H_i^M = \sum_{j=1}^{n} (P_{ij} / \sum_{j=1}^{n} P_{ij}).(P_{ij} / \sum_{i=1}^{m} P_{ij})$$

where S_i is the sensitivity, which measures the extent to which proteins belonging to one MIPS complex are grouped within the same complex defined here, and H^M_i is the homogeneity, measuring the extent to which proteins from the same MIPS complexes are distributed across our complexes (clusters). Similarly, for each of the C_j complexes we computed the Positive Predictive Value PPV_j, which measures the fraction of components of a cluster which belong to the same MIPS complex, and represents thus the reliability with which the cluster ‘predicts’ this complex, and the homogeneity $\text{H}^{\text{C}}{}_{\text{j}}$, evaluating the extent to which proteins from one cluster are distributed among different MIPS complexes (Brohée et al, unpublished)

$$PPV_j = \max_i(P_{ij}) / \sum_{i=1}^{m} P_{ij} \quad ; \qquad H_j^C = \sum_{i=1}^{m} (P_{ij} / \sum_{i=1}^{m} P_{ij}).(P_{ij} / \sum_{j=1}^{n} P_{ij})$$

To evaluate the overall correspondence between the two sets of complexes, we compute the S_{mean} and PPV_{mean} as the weighted means of S_{i} and PPV_{j} across columns and rows, respectively, as well as the means of each of the homogeneity scores $\text{H}^{\text{M}}{}_{\text{i}}$ and $\text{H}^{\text{C}}{}_{\text{j}}$. The overall correspondence is then given by two scores computed as the geometric means of the corresponding means defined above:

$$\text{Precision}_{\text{tot}} = \text{sqrt}(\text{S}_{\text{mean}} * \text{PPV}_{\text{mean}})$$

$$\text{Homogeneity}_{\text{tot}} = \text{sqrt } (\text{H}^{\text{M}}{}_{\text{mean}} * \text{H}^{\text{C}}{}_{\text{mean)}}$$

2-Diversity of the Go Process annotations in genes within complexes (Figure S2)

To quantify functional similarity between pairs of proteins, we apply a semantic similarity measure[7] to the Gene Ontology (GO) terms with which these proteins are annotated [GO, 2000]. The semantic similarity measure takes into account the relative frequency and level of hierarchy of GO terms in the 'Biological process' taxonomy.

In this analysis the semantic similarity was evaluated for all protein pairs within each complex, averaged over all pairs in each complex and over all complexes. In addition we also computed the average semantic similarity score per interaction within complexes.

3-Diversity of the cellular localizations for genes within complexes (Figure S1)

To evaluate the extent to which proteins in the same cluster/complex have the same cellular localization we used the cellular localizations determined experimentally[8] . The different localization categories were treated as groups into which proteins in our complexes were assigned and the fraction of the proteins in each complex/cluster *j* that map into the same localization category *i* was computed as the per cluster Positive Predictive Value $PPV_j = \max_i(P_{ij}) / \sum_{i=1}^{m} P_{ij}$, where *Pij* are the proteins in cluster *j* assigned to the experimental localization category *i*. The weighted average of the PPV_j values is then computed to yield the PPV*tot* (computed as described above for the overlap with the MIPS complexes), which is used as the global measure for how well proteins in the derived complexes are co-localized over the entire dataset.

References

1. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
2. Mewes, H. W. et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32**, D41-4 (2004).
3. King, A. D., Przulj, N. & Jurisica, I. Protein complex prediction via cost-based clustering. *Bioinformatics* **20**, 3013-20 (2004).
4. Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* **100**, 12123-8 (2003).
5. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
6. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-84 (2002).
7. Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 1275-83 (2003).
8. Huh et al.Global analysis of protein localization in budding yeast, *Nature* **425**, 686-691(2003).

Supplementary Information List of Tables and Figures

Table S1 – List of all proteins whose purification was attempted
Table S2 – List of successful purifications
Table S3 - List of all proteins identified

Table S4 – List of background and contaminant proteins removed
Table S5 – Intersection dataset
Table S6 – Merged dataset
Table S7 - Core dataset interactions
Table S8 – Extended Core dataset interactions
Table S9 – Complete set of interactions from machine learning
Table S10 - List of protein complexes

Figure S1 – Co-localization of MIPS, Gavin, Ho, Core, Extended Core datasets
Figure S2 – Semantic similarity (GO biological processes) for all
Figure S3 – Cytoscape view indicating comparison with MIPS
Figure S4 - Essentiality versus conservation, degree of connectivity and betweenness
Figure S5 – IWR1 complex data
Figure S6—Screenshot of the yeast interactome database

Figure legends for supplementary data.

Figure S1. Positive prediction values (Sylvain Brohee, manuscript in preparation) for large-scale datasets using subcellular localization data from Huh et al., 2003 Nature 425, 671.

Figure S2. Semantic similarity scores (Lord et al. 2003 Bioinformatics 19, 1275) for large-scale datasets using GO biological process taxonomy (www.geneontology.org)

Figure S3. Screenshot of Cytoscape/GenePro representation of Core protein complex network. Each node represents an individual complex. An individual complex (Predicted complex #50) is enlarged in inset. Red indicates members of the complex while blue indicates neighbours of degree one.

Figure S4. Essential genes are more conserved, connected and have more betweenness than non-essential genes

Figure S5. A. Summary of purifications using tagged Iwr1 and tagged unique subunits of RNA polymerase II. B. Amino acid alignment of Iwr1 (Ydl115c) sequence from various species.

Figure S6. A comprehensive database of the yeast interactome (http://tap.med.utoronto.ca/). Querying of a protein name (**1**) leads to an interaction view summarizing the results of all purifications involving the queried protein, as well as links to YPD, SGD, MIPS and BioGRID databases. Links from this table lead to details of the LC/MS/MS (**2**) or MALDI-TOF (**6** and **7**) experiments. From the LC/MS/MS page, mass spectrum search algorithm scores (**2** and **3**) and spectral images (**5**) can be retrieved, in addition to results from SIMS site-specific modification searches (**4**). Similarly, the link to the MALDI-TOF data allows for retrieval of search algorithm scores (**7**), silver-stained gels (**6**) and spectral images (**8**).

Table legends for supplementary data

Supplementary Table 1. List using systematic nomenclature of all the 4562 proteins of *S. cerevisiae* that were successfully tagged and whose purification was attempted.

Supplementary Table 2. List using systematic nomenclature of all 2357 *S. cerevisiae* proteins whose purification was successful. For the purposes of the paper and this table, success was defined as the recovery and high-confidence identification of at least one non-background protein. This allowed us to compensate for the minority of cases in which an interacting protein(s) but not the bait was successfully identified by mass spectrometry.

Supplementary Table 3. List of 4087 proteins that were identified during the course of this project. Of these, 4021 were identified with high confidence (>99%) by MALDI-TOF mass spectrometry and/or LC/MS/MS, and an additional 66 were identified with lower confidence (<99%) by MS but were identified as a partner in a high confidence interaction (i.e. they were identified repeatedly by MS as interacting proteins with lower MS confidence, so that the machine learning algorithms identified them anyway as a high confidence partners). The 66 proteins in this list identified only as interaction partners rather than with at least 99% confidence by mass spectrometry were YIL044C: AGE2, YBR132C: AGP2, YHR129C: ARP1, YPR020W: ATP20, YBR003W: COQ1, YHR055C: CUP1-2, YOR180C: DCI1, YNL001W: DOM34, YNL133C: FYV6, YGL254W: FZF1,YDR302: GPI11, YIR038C: GTT1, YOR278:W HEM4, YFL011W: HXT10, YNL078W: JIP1, YBR015C: MNN2, YOL042:W NGL1, YOR104W: PIN2, YMR006C: PLB2, YFR033C: QCR6, YER173W: RAD24,YDR014W: RAD61, YLR453C: RIF2, YHL027W: RIM101, YMR305C: SCW10, YLL011W: SOF1, YGL169W: SUA5, YDR007W: TRP1, YDL064W: UBC9, YPL252C: YAH1, YBR042C, YBR053C, YBR111C, YCR023C, YDL085C-A, YDR061W, YDR063W, YDR262W , YDR320C-A, YDR458C, YER048W-A, YER134C, YFR018C, YGR201C, YGR205W, YHR046C, YHR198C, YIL055C, YIL157C, YJL062W-A, YJR088C, YKL005C, YKR074W, YLR285W, YLR363W-A, YLR376C, YMR290W-A, YNR034W-A, YNR063W, YOL032W, YOR285, YPR148C, YPR172W, YBL071W, YFR024C, and YLR438C.

Supplementary Table 4. List of *S. cerevisiae* proteins, including a number of chaperones, that were identified by MALDI-TOF MS or LC/MS/MS in more than 3% (71) of all the successful protein purifications. For the purposes of identifying protein-protein interactions these proteins were considered non-specific contaminants and were removed from consideration. The listed cytoplasmic ribosomal proteins were also removed. Because of this, interactions of ribosomal proteins with each other in the ribosome or with many translation factors are not present in our protein interaction network. As well, some chaperones have interactions with many proteins in the normal course of their functioning but were removed from our lists of protein-protein interactions.

Supplementary Table 5. List of 2357 protein-protein interactions involving 1210 proteins in the "intersection dataset," along with the confidence scores for protein identification by mass spectrometry. An interaction is listed in this Table if the interacting protein partner was identified in a single purification by both MALDI-TOF mass spectrometry and LC/MS/MS, and even if the confidence for protein identification by mass spectrometry was relatively low, as long as it was greater than 70%. As described in Supplementary Information, the Z-scores for protein identification by mass spectrometry were rescaled to a scale from 0 to 1 such that a score of 0.5 represents 70% confidence.

Supplementary Table 6. List of 5496 protein-protein interactions involving 2186 proteins in the "merged dataset," along with the confidence scores for protein identification by mass spectrometry. This list contains all 2357 interactions in the "intersection dataset," as well as interactions identified repeatedly by only a single method of mass spectrometry (with a confidence threshold of 70%) and interactions identified by a single mass spectrometry method (with a confidence threshold of 70%) when each of the interacting protein partners was purified. As described in Supplementary Information, the Z-scores for protein identification by mass spectrometry were rescaled to a scale from 0 to 1 such that a score of 0.5 represents 70% confidence.

Supplementary Table 7. List of 7123 protein-protein interactions in the core dataset involving 2708 *S. cerevisiae* proteins. These were the highest confidence interactions (minimum 0.273, mean 0.68, median 0.69) identified by the machine learning procedure described in the main text and Figure 2a. Each interaction is listed along with its confidence score. It is these protein-protein interactions that were used to compute the protein interaction network and identify protein complexes as described in the text and shown in Figures 3d and S3.

Supplementary Table 8. List of 14317 protein-protein interactions in the extended dataset involving 3672 *S. cerevisiae* proteins. Each interaction is listed together with its confidence score as determined by the machine learning procedure. This list includes all the protein-protein interactions listed in Table S8 but now the minimum confidence score is 0.101 rather than 0.273. This Table should have approximately 1000 correct interactions among the 7194 interactions that are not listed in Table S8.

Supplementary Table 9. Complete list of all the putative *S. cerevisiae* protein-protein interactions identified in this study, together with their confidence scores as determined by the machine learning algorithms. The vast majority of the interactions listed here that are not listed in Table S9 are incorrect. They have very low confidence scores (usually much less than 0.1) either because the interactions were not identified reproducibly or because the protein identifications by mass spectrometry had low confidence, or both. This list is provided because some of the additional interactions listed here and not in Tables S8 and S9, perhaps as many as several hundred, could be correct.

Supplementary Table 10. List of protein complexes and their component subunits as identified by the Markov Cluster Algorithm. Each of the protein clusters is given a number from 0 to 546. The MCL algorithm does not do well at separating protein complexes that share subunits. At the moment, achieving this aim for a given protein complex is likely to require manual inspection of the data.