Supplementary Information

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

Networks built in this study from the experimentally observed interactions:

<table>
<thead>
<tr>
<th>Network</th>
<th>#genes/nodes</th>
<th># interactions/edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extended Core Network*</td>
<td>3672</td>
<td>14317</td>
</tr>
<tr>
<td>Core Network*</td>
<td>2708</td>
<td>7123</td>
</tr>
<tr>
<td>Merged Network (R-)</td>
<td>2186</td>
<td>5496</td>
</tr>
<tr>
<td>Intersection Network (R-)</td>
<td>1210</td>
<td>2357</td>
</tr>
</tbody>
</table>

Extended core network includes all interactions with scores ≥0.101.
Core network is the same network as above but includes only those interactions with scores ≥0.273. 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.

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 a network; to visually integrate the 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 set of ‘plug-ins’ that provide several integrative and interactive visualization and analysis capabilities for networks of interacting proteins and genes to be described elsewhere (Orsi et al., submitted).

The highly connected modules identified by the MCL clustering procedure, which we take here to represent multi-protein complexes characterized in this study, are displayed and analyzed using the GenePro Cytoscape plug-ins. 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 the another 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 indicating 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 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 color 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. 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-ins enable detailed analysis of proteins and interactions within individual complexes. A left Mouse click on a complex node displays a new network, which represents all the proteins in the complex as nodes of one color (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 neighbors in the protein-protein interaction network. These neighbors 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 color from the proteins within the complex (blue in Fig. S3). All displayed intra- and inter-complex edges are colored 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 to
query 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.

**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 \(C_1\ldots C_n\) complexes/clusters computed in this work, and the \(M_1\ldots M_m\) 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 Mᵢ 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 Mᵢ we then compute 2 quantities (Brohée et al., unpublished data):

\[ Sᵢ = \max_j \left( \frac{\sum_{j=1}^{n} Pᵢₗ}{Pᵢ} \right) \]
\[ Hᵢ^M = \sum_{j=1}^{n} \left( \frac{\sum_{j=1}^{n} Pᵢₗ}{\sum_{j=1}^{n} Pᵢₗ} \right) \left( \frac{\sum_{i=1}^{m} Pᵢₗ}{\sum_{i=1}^{m} Pᵢₗ} \right) \]

where \( Sᵢ \) is the sensitivity, which measures the extent to which proteins belonging to one MIPS complex are grouped within the same complex defined here, and \( Hᵢ^M \) is the homogeneity, measuring the extent to which proteins from the same MIPS complexes are distributed across our complexes. Similarly, for each of the Cⱼ complexes we computed Positive Predictive Value PPVⱼ, 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 \( Hⱼ^C \), evaluating the extent to which proteins from one cluster are distributed among different complexes (Brohée et al, to be published)

\[ PPVⱼ = \max_i \left( \frac{\sum_{i=1}^{m} Pᵢₗ}{\sum_{i=1}^{m} Pᵢₗ} \right) \]
\[ Hⱼ^C = \sum_{i=1}^{m} \left( \frac{\sum_{i=1}^{m} Pᵢₗ}{\sum_{i=1}^{m} Pᵢₗ} \right) \left( \frac{\sum_{j=1}^{n} Pᵢₗ}{\sum_{j=1}^{n} Pᵢₗ} \right) \]

To evaluate the overall correspondence between the two sets of complexes, we compute the \( S_{\text{mean}} \) and \( PPV_{\text{mean}} \) as the weighted means of \( Sᵢ \) and \( PPVⱼ \) across columns and rows, respectively, as well as the means of each of the homogeneity scores \( Hᵢ^M \) and \( Hⱼ^C \). The overall correspondence is then given by two scores computed as the geometric means of the corresponding means defined above:
\[
\text{Precision}_{\text{tot}} = \sqrt{S_{\text{mean}} \times \text{PPV}_{\text{mean}}}
\]

\[
\text{Homogeneity}_{\text{tot}} = \sqrt{H_{\text{mean}}^M \times H_{\text{mean}}^C}
\]

2-Diversity of the Go Process annotations in genes within complexes

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

To evaluate the extent to which proteins in the same cluster/complex have the same cellular localization we used the cellular localizations determined experimentally (Huh et al.). 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 \(P_{ij}\) 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 \(PPV_{\text{tot}}\) (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

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 post-translational modifications
Table S5 – List of background and contaminant proteins removed
Table S6 – Intersection dataset
Table S7 – Merged dataset
Table S8 - Core dataset interactions
Table S9 – Extended Core dataset interactions
Table S10 – Complete set of interactions from machine learning
Table S11 - 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 legends for supplementary data.

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

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 lwr1 and tagged unique subunits of RNA Polymerase II. B. Amino acid alignment of lwr1 (Ydl115c) sequence from various species.