Large−scale detection of protein−protein interactions: a comparative assessment

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P

Supplementary Information

1) Datasources

Yeast two−hybrid interactions
957 interactions were taken from Uetz et al[1], 4549 interactions from Ito et al[2]. This yielded a total of 5127 high−throughput yeast two−hybrid interactions involving 3579 proteins (overlapping interactions were counted only once, homotypic interactions were not counted).

Purified complexes
The two datasets based on systematic purification of protein complexes[3,4] are the largest interaction datasets to date, which is why we considered them separately. We focussed on the filtered datasets (both groups removed 'sticky' proteins and components of the ribosome), and assigned connections between all proteins present in a purification. Apart from filtering, we considered only the raw data before any manual curation or complex assignment. The data are available as supplementary material from the respective publications, and for the HMS−PCI data also from http://www.mdsp.com/yeast/MDSP−Nature10Jan02−YeastComplexes.txt
In the HMS−PCI approach, the same bait was occasionally purified more than once, often under different conditions (e.g. with or without DNA−damaging agents). We chose to collapse such repeated purifications of the same bait into one experiment. We verified that this approach has no major influence on the results present here (When keeping these purifications separate, the number of interactions goes down from 33014 to 31048. However, the coverage of known interactions also goes down somewhat from 668 to 665 interactions, as does the contribution to the overlap from 1997 to 1971 interactions.)

In silico data
This dataset contains contributions from three different methods.
A) conserved gene neighborhood[5,6] − we searched 42 completely sequenced genomes for instances of conserved neighborhood between genes. We required two or more genes to have the same orientation on the chromosome and to be in a 'run' with intergenic regions of no more than 300 bp[5]. Additionally, we required corresponding evidence (for orthologous genes) in at least one other, diverged organism.
B) co−occurrence of genes[7,8] − For each entry in the orthology−database COG[9], we recorded the pattern of occurrence among 42 completely sequenced genomes. We compared these patterns and recorded a putative functional interaction between those whose mutual information[10] was higher than 0.5 (close matches to the 13 most frequent patterns were ignored, as they are mostly phylogenetic).
C) Gene fusion events − we detected gene fusions by the presence of a gene in more than one COG cluster. Single fusion events were considered significant.
The methods outlined above yield interaction predictions between orthologous groups of genes, not individual genes. We mapped this to individual yeast genes using the COG database, by expanding links between two COG clusters to links between all yeast genes in these clusters. Some COG entries contain a large number of yeast−proteins (gene families expanded in eukaryotes), so a cutoff was chosen in order not to allow more than six links between similar yeast genes for two COG clusters.
A total of 7446 interaction predictions are contained in the *in silico* dataset (6387 from conserved gene neighborhood, 358 from gene fusions, 997 from co−occurence of genes, the predictions

overlap). For gene fusions and co−occurence, parameter choice was less inclusive than in ref 11, which is reflected in a much lower coverage. However, this leads to a somewhat higher accuracy.

correlated mRNA expression
This dataset is based on two large, genome−wide surveys of mRNA expression in yeast. One is a compendium of gene expression under 300 different cellular conditions (ref 12, diverse mutations and chemical treatments), the other is a study of the mitotic cell cycle[13], in which the status of gene expression is measured at 17 different time−points in synchronized cultures of yeast.
All mRNA levels were converted to log−ratios (natural logarithm of the ratio of measured expression by reference expression), and subjected to z−score normalization (which ensures that the mean of all values is zero and the standard deviation one). We fused both datasets, yielding 317 measurements per gene. For all possible pairs of genes, we then computed the Pearson correlation coefficient to measure the similarity of their expression profiles. All pairs having a similarity above a given cutoff were connected by a putative interaction. The cutoff was chosen to provide a compromise between coverage and accuracy (in figure 2, several different cutoffs are plotted. For figure 1, the cutoff was 0.675).

Genetic interactions
295 synthetic lethal interactions from the first high−throughput study on genetic interactions in yeast[14], plus an additional 591 synthetic lethal interactions parsed from the MIPS database (ref 15, see http://mips.gsf.de/proj/yeast/tables/interaction/genetic_interact.html)

2) Functional categories (ad Figure 1)

We assigned each yeast ORF to one of 12 broad functional categories (or to the category 'uncharacterized') using the hierarchical classification of gene function performed at MIPS[15] as a template (see http://mips.gsf.de/proj/yeast/catalogues/funcat/index.html). Genes that were annotated in more than one category were manually placed into just one. Some MIPS categories were fused for conciseness.

The following MIPS−categories were used:

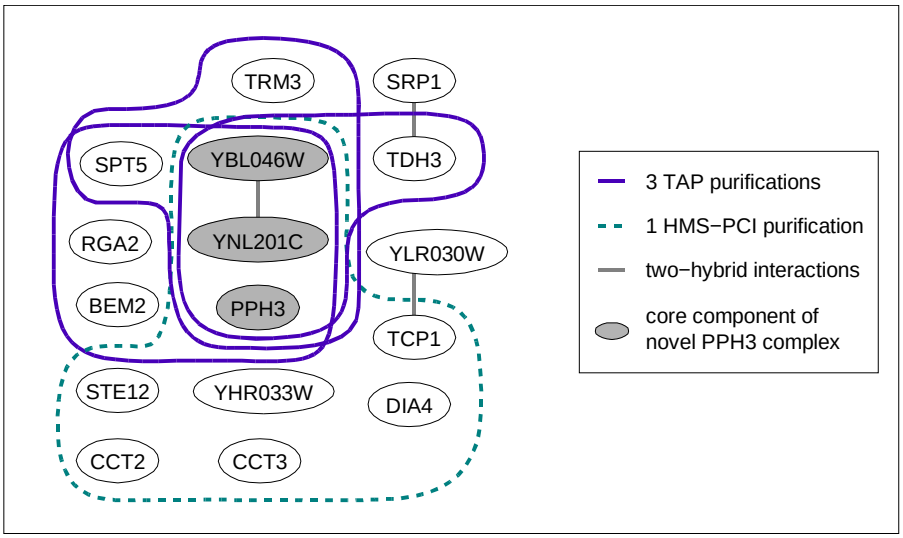
category	Description	Original MIPS category
E	energy production	energy
G	aminoacid metabolism	aminoacid metabolism
M	other metabolism	all remaining metabolism categories
P	Translation	protein synthesis
T	Transcription	transcription, but without subcategory 'transcriptional control'
B	transcriptional control	subcategory 'transcriptional control'
F	protein fate	protein fate (folding, modification, destination)
O	cellular organization	cellular transport and transport mechanisms
A	transport and sensing	categories 'transport facilitation' and 'regulation of / interaction with cellular environment'
R	stress and defense	cell rescue, defense and virulence
D	genome maintenance	DNA processing and cell cycle
C	cellular fate / organization	categories 'cell fate' and 'cellular communication / signal transduction' and 'control of cellular organization
U	Uncharacterized	categories 'not yet clear−cut' and 'uncharacterized'

For the full list of category assignments, see Table S1.

3) Glycine decarboxylase

The glycine decarboxylase complex is a multienzyme complex needed when glycine is used as a one−carbon source. It is under tight transcriptional control, and the key components GCV1, GCV2, and GCV3 are induced only when there is excess glycine and cytoplasmic 5,10−CH2−H4−folate levels are low[16] . This is presumably not the case under the conditions used for the large−scale purification of complexes, which may be an explanation why the complex is not detected (GCV3 was TAP−tagged and failed to present other members of the complex).
The three components can be confidently linked, however, using three different lines of evidence: a) All three genes repeatedly occur next to each other in prokaryotic genomes, forming an operon in seven diverged species. b) they show a very similar phylogenetic distribution: whenever one of the genes is missing from a genome, the other two are also absent. This is the case for a total of 17 genomes, many from organisms with a parasitic lifestyle, but also from three archaea living under extreme conditions. c) microarray experiments in yeast confirm that the three genes are closely co−regulated transcriptionally, even though they are not in close neighborhood to each other on the genome in yeast.

4) The PPH3 protein



The PPH3 protein is a serine/threonine phosphatase related to the PP2A family of phosphatases[17,18]. There are several such phosphatases in the yeast genome; very little is known about PPH3. Through high−throughput interaction data, this phosphatase can be confidently linked to two proteins of unknown function, YBL046W and YNL201C. It is found together with them in four independent purifications, and the two unknown proteins are also linked through one two−hybrid interaction (only the Ito−core set is shown). Other proteins are also found linked to this core, however not consistently. It is unclear whether these are false or true positives, and some are themselves linked to yet other proteins (not shown). From high−throughput data alone, it is difficult to say whether the complex consists of three or more than three proteins.

5) Reference set / benchmark (ad Figure 2)

We assembled a reference set of known interactions from two catalogs of protein complexes in yeast. One (http://mips.gsf.de/proj/yeast/catalogues/complexes/index.html) is maintained at MIPS, the other can be accessed free of charge for academic users at http://www.incyte.com/ or

http://www.proteome.com/ (Bioknowledge database − YPD[19]). We assigned binary interactions between all proteins participating in a complex. Where complex annotation was hierarchical, we stayed at the lowest level of hierarchy, i.e. we only considered subcomplexes and not larger assemblies. 8852 interactions (involving 999 proteins) were derived from the MIPS complexes, and we complemented these with 2055 additional interactions (involving 309 proteins) from YPD − a manually retrieved subset of their interaction data − bringing the total to 10907 interactions involving 1308 proteins.
For the benchmark, we compared binary interactions contained in the individual datasets against the binary interactions in the reference set. For the purified complexes, this meant assigning binary interactions between all members of a complex.

As an alternative, we also followed the approach taken in the HMS−PCI study − to assign only interactions between the bait and the associated proteins, but not among the associated proteins. We find that this leads to higher accuracy−values for both the HMS−PCI and the TAP datasets, but concomitantly to a lower coverage, see below.
Additionally, we repeated the calculations from yet another angle − we considered only interactions where proteins A and B are actually present in the reference set. This leads to still higher accuracy values, but fails to take into consideration more then 60% of the available data for each dataset.

dataset (interactions)	Accuracy	Coverage
TAP (all interactions)	12.5 %	21 %
TAP (bait interactions only)	27.8 %	8.5 %
TAP (within reference set only)	40.5 %	not applicable
HMS−PCI (all interactions)	2.0 %	6.1 %
HMS−PCI (bait interactions only)	6.8 %	2.3 %
HMS−PCI (within reference set only)	14.2 %	not applicable
Yeast two hybrid (all interactions)	3.7 %	1.7 %
Yeast two hybrid (within reference only)	38.1%	not applicable

6) Randomizing datasets (significance of overlap)

We repeatedly randomized the high−throughput interactions and the reference set: For each set, the exact number of binary interactions was maintained, but the two proteins forming a binary interaction were randomly chosen from the complete set of yeast proteins. We assayed how many of these interactions were overlapping (detected by more than one high−throughput method). In 15 independent randomizations, only 118.4 interactions were on average overlapping (compared to 2455 overlapping interactions in the real data).

7) Filtering (ad Figure 2)

For some datasets, we show in Figure 2 raw and filtered data. For the yeast two−hybrid data, the raw set is the combination of all interactions detected[1,2]. The filtered set is the 'core−data' of ref. [2]. For the purified complexes based on the HMS−PCI approach, the raw data is the data in Supplementary Table S1 of ref [4] (note that this table does not contain any ribosomal proteins). The filtered set is the data from Supplementary Table S2 (which is derived from S1 by removing unspecific interactors). For the complexes based on the TAP−tag, the raw data is the purifications

before removal of any unspecific interactors or ribosomal proteins. The filtered data is the data contained in Supplementary Table S1 of ref [3]. For Figure 2, we verified that the difference between the HMS−PCI and TAP approaches is not simply caused by differences in bait selection: When considering only those purifications where both approaches tested the same bait, the numbers for accuracy (TAP vs HMS: 15.2% vs 2.9%) and coverage (6.1% versus 2.5%) show the same trend as for the full datasets.
For the insilico data, the set with the lowest coverage is the one used for Figure 1. For the medium coverage, we allowed for more redundancy in mapping orthologues between prokaryotes and yeast − a maximum of 12 interactions between yeast−genes (instead of six) was allowed for each pair of interacting orthologous clusters. Additionally, the mutual information cutoff for the cooccurence interactions was lowered to 0.4125, while still filtering out the eight most frequent phylogenetic patterns. Finally, for the third dataset (highest coverage but lowest accuracy), we further lowered the stringency for gene−neighborhood interactions: normally, two instances of gene neighborhood in diverged species are required. Here, we counted also single instances, but only if the two orthologous groups were not too large (cutoff was 1060 for the product of their sizes).
For the interactions predicted by correlated mRNA expression (synexpression), interactions were computed as in Figure 1, but various pearson correlation coefficients were used as cutoffs: they range from 0.6 (lowest coverage) to 0.35 (best coverage), in steps of 0.05

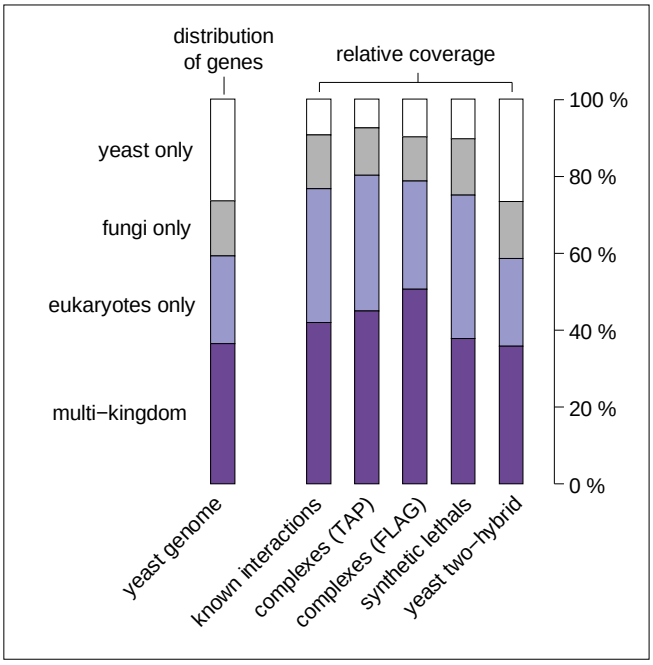
8) mRNA abundance (ad Figure 3)

We relied on a genome−wide analysis of mRNA abundances in yeast[20]. The raw data of this analysis can be downloaded here: http://web.wi.mit.edu/young/expression/transcriptome.html. We used this data to separate the yeast genome into 10 bins of equal size, sorted according to mRNA abundance (for about 10% of the yeast genome, mRNA−abundance data is not available from ref 20. These genes are most likely not expressed at all or were not measured). For each interaction dataset and each bin, we then counted how many interactions had at least one partner belonging to the bin. Here, each binary interaction is counted twice − once for partner A, and once for partner B. We do not show the data for correlated mRNA expression, because we found the result to be strongly dependent upon the similarity measure chosen for analysis. When choosing euklidian distance as a similarity measure, correlated mRNA showed a bias for strongly expressed genes, when using the pearson correlation coefficient, it did not.

9) Protein localizations (ad Figure 4)

We derived protein localizations from the MIPS[15] and TRIPLES[21] databases. In cases where both databases have an entry for a protein, we prioritize MIPS (it is based on literature information). We parsed 2342 protein localizations from MIPS, and an additional 1511 from TRIPLES.

10) A bias towards well−conserved proteins



We separated the yeast genome into four classes according to the conservation of the genes in other species. To check for conservation in prokaryotes ("multi kingdom"), we used the COG database. The other three classes are:
a) eukaryotes − H. sapiens, D. melanogaster, C. elegans and A. thaliana
b) Fungi only − S. pombe and C. albicans
c) yeast only.
The presence of a gene in any of these species was concluded from bi−directional best hits in Swiss−Waterman searches, using 0.01 as a cutoff.
The resulting classification of the yeast genome is available in Table S2.

11) A detailed look at the overlap between TAP and HMS−PCI

When analyzing the overlap between the TAP and HMS−PCI approaches, limiting the analysis to the shared baits is insufficient. Similar complexes are often detected by both methods, even when using different baits as entry−points. This ’reverse tagging’ complementarity of the individual purifications is a general advantage of the system. Check Table S3 for a sorted list of overlapping purifications. The largest overlap between two purifications is 18 shared proteins.
Other numbers characterizing the overlap:
1379 proteins are touched by TAP (filtered set)
1578 proteins are touched by HMS−PCI (filtered set)
659 proteins are touched by both. Note that the TAP approach specifically lists proteins that do not detect any interactions (singletons), whereas HMS−PCI filters these from their dataset.
among the shared proteins, HMS−PCI detects 9005 connections, TAP 6285.
the overlap is 1728 interactions, i.e. 27.5% of the TAP data overlap with 19.2% of the HMS−PCI data.

When measuring how similar the purifications are in terms of the number of shared proteins, the overlap is more difficult to define. Which purifications to compare with which?
We chose an approach in which we, for each purification, determined the ’best matching’ purification in the other dataset (considering only those that shared at least one protein. 404 TAP purifications share at least one protein with the HMS−PCI purifications, and 399 HMS−PCI purifications share at least one protein with the TAP purifications). Among those, on average 38.8 % of the TAP data overlap with 31.6 % of the HMS−PCI data. When only comparing those purifications that share a bait (of which there are 94), the overlap stands at 42.4 % of the TAP data and 33.0 % of the HMS−PCI data. This is when including the baits as shared proteins (overlap) − when counting only proteins other than the baits, the numbers are 29.3 % and 18.0 %, respectively.

12) TAP/HMS−PCI: technical differences

HMS−PCI approach[4]

+ high analytical depth through use of LC−MS/MS on every band.
+ one−step purification: more transient interactions.
+ can get more baits to work.
+ can overexpress to detectable amounts if needed.
+ short term induction of expression: good if modified bait is toxic.
+ expression from plasmids: not limited to yeast proteins.
− artifacts due to overexpression (stoichiometry may be changed).
− one−step purification: more background.
− slow and more expensive MS approach.
− untagged protein still present in the genome: possible competition for binding partners

TAP approach[3]

+ endogenous promoter: wild−type expression level, stoichiometry of binding partners largely maintained.
+ two−step purification, very mild conditions: high sensitivity and specificity of purifications.
+ conditions more homogenous (no specific induction necessary).
+ fast and sensitive MS approach (MALDI−TOF, can detect proteins at 15 copies per cell).
+ genomic integration: no untagged version of protein present.
− two−step purification: longer, may wash off transient binding partners.
− endogenous promoter: failure if expression is below certain threshold.
− technical bias against proteins below 10 kD (MS, tagging, etc...)

13) Combination of high−throughput data

Table S4 contains a compilation of all high−throughput interactions studied here. Binary interactions are shown, sorted according to confidence − high confidence interactions are shown first.

References (for the supplementary material, this text)

1. Uetz, P. et al. A comprehensive analysis of protein−protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623−7 (2000).
2. Ito, T. et al. A comprehensive two−hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569−74 (2001).
3. Gavin, A. C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141−147 (2002).
4. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180−183 (2002).
5. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896−901 (1999).
6. Huynen, M., Snel, B., Lathe, W., 3rd & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10**, 1204−10 (2000).
7. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285−8 (1999).

8. Huynen, M. A. & Bork, P. Measuring genome evolution. *Proc Natl Acad Sci U S A* **95**, 5849−56 (1998).
9. Tatusov, R. L. et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22−8 (2001).
10. Korber, B. T., Farber, R. M., Wolpert, D. H. & Lapedes, A. S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A* **90**, 7176−80 (1993).
11. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome−wide prediction of protein function. *Nature* **402**, 83−6 (1999).
12. Hughes, T. R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109−26 (2000).
13. Cho, R. J. et al. A genome−wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65−73 (1998).
14. Tong, A. H. et al. Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science* **294**, 2364−2368 (2001).
15. Mewes, H. W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31−4 (2002).
16. Piper, M. D., Hong, S. P., Ball, G. E. & Dawes, I. W. Regulation of the balance of one−carbon metabolism in Saccharomyces cerevisiae. *J Biol Chem* **275**, 30987−95 (2000).
17. Hoffmann, R., Jung, S., Ehrmann, M. & Hofer, H. W. The Saccharomyces cerevisiae gene PPH3 encodes a protein phosphatase with properties different from PPX, PP1 and PP2A. *Yeast* **10**, 567−78 (1994).
18. Kalhor, H. R., Luk, K., Ramos, A., Zobel−Thropp, P. & Clarke, S. Protein phosphatase methyltransferase 1 (Ppm1p) is the sole activity responsible for modification of the major forms of protein phosphatase 2A in yeast. *Arch Biochem Biophys* **395**, 239−45 (2001).
19. Costanzo, M. C. et al. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* **29**, 75−9 (2001).
20. Holstege, F. C. et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717−28 (1998).
21. Kumar, A. et al. The TRIPLES database: a community resource for yeast molecular biology. *Nucleic Acids Res* **30**, 73−5 (2002).