Sciencexpress Report

High-Quality Binary Protein Interaction Map of the Yeast Interactome Network

Haiyuan Yu,[1,2*] Pascal Braun,[1,2*] Muhammed A. Yildirim,[1,2,3*] Irma Lemmens,[4] Kavitha Venkatesan,[1,2] Julie Sahalie,[1,2] Tomoko Hirozane-Kishikawa,[1,2] Fana Gebreab,[1,2] Na Li,[1,2] Nicolas Simonis,[1,2] Tong Hao,[1,2] Jean-François Rual,[1,2] Amélie Dricot,[1,2] Alexei Vazquez,[5] Ryan R. Murray,[1,2] Christophe Simon,[1,2] Leah Tardivo,[1,2] Stanley Tam,[1,2] Nenad Svrzikapa,[1,2] Changyu Fan,[1,2] Anne-Sophie de Smet,[4] Adriana Motyl,[6] Michael E. Hudson,[6] Juyong Park,[1,7] Xiaofeng Xin,[8] Michael E. Cusick,[1,2] Troy Moore,[9] Charlie Boone,[8] Michael Snyder,[6] Frederick P. Roth[1,10], Albert-László Barabási[1,7], Jan Tavernier,[4] David E. Hill,[1,2] Marc Vidal[1,2†]

[1]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA 02115, USA. [2]Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. [4]Department of Medical Protein Research, VIB, and Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, 9000 Ghent, Belgium. [5]The Simons Center for Systems Biology, Institute for Advanced Studies, Princeton, NJ 08540, USA. [6]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06620, USA. [7]Center for Complex Network Research and Departments of Physics, Biology and Computer Science, Northeastern University, Boston, MA 02115, USA. [8]Banting and Best Department of Medical Research and Department of Medical Genetics and Microbiology, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada M5S 3E1. [9]Open Biosystems, Huntsville, AL 35806, USA. [10]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work.
[†]To whom correspondence should be addressed. E-mail: marc_vidal@dfci.harvard.edu

Current yeast interactome network maps contain several hundred molecular complexes with limited and somewhat controversial representation of direct binary interactions. We carried out a comparative quality assessment of current yeast interactome datasets, demonstrating that high-throughput yeast two-hybrid (Y2H) provides high-quality binary interaction information. As a large fraction of the yeast binary interactome remains to be mapped, we developed an empirically-controlled mapping framework to produce a "second-generation" high-quality high-throughput Y2H dataset covering ~20% of all yeast binary interactions. Both Y2H and affinity-purification followed by mass spectrometry (AP/MS) data are of equally high quality but of a fundamentally different and complementary nature resulting in networks with different topological and biological properties. Compared to co-complex interactome models, this binary map is enriched for transient signaling interactions and inter-complex connections with a highly significant clustering between essential proteins. Rather than correlating with essentiality, protein connectivity correlates with genetic pleiotropy.

A crucial step towards understanding cellular systems properties is mapping networks of physical DNA-, RNA- and protein-protein interactions, the "interactome network", of an organism of interest as completely and accurately as possible. One approach consists in systematically testing all pairwise combinations of predicted proteins to derive the "binary" interactome. Early attempts at binary interactome mapping used high-throughput yeast two-hybrid (Y2H), in which a protein interaction reconstitutes a transcription factor that activates expression of reporter genes. High-throughput Y2H maps have been generated for *Saccharomyces cerevisiae* (*1–3*), *Caenorhabditis elegans* (*4–6*), *Drosophila melanogaster* (*7*), and human (*8–10*). An alternative approach consists in generating "co-complex" interactome maps, achievable by high-throughput co-affinity purification followed by mass spectrometry (AP/MS) to identify proteins bound to tagged baits, as done for *Escherichia coli* (*11, 12*), *S. cerevisiae* (*13–16*), and human (*17*).

To investigate fundamental questions of interactome network structure and function, it is necessary to understand how the size and quality of currently available maps, including thorough evaluation of differences between binary

and co-complex maps, might have affected conclusions about global and local properties of interactome networks (*18, 19*). Here, we address these issues using the yeast *S. cerevisiae* as a model system.

First, we compared the quality of existing high-throughput binary and co-complex datasets to information obtained from curating low-throughput experiments described in the literature (Fig. 1A). For binary interactions we examined: (i) the subset found by Uetz *et al.* in a proteome-scale all-by-all screen ("Uetz-screen"), excluding the pairs found in a focused, potentially biased experiment involving only 193 baits ("Uetz-array") (*2*); and (ii) the Ito *et al.* interactions found three times or more ("Ito-core"), independently from those found one or two times ("Ito-noncore"), a distinction recommended by the authors but seldom applied in the literature (*3*). For co-complex associations, we investigated two high-throughput AP/MS datasets referred to as "Gavin" (*15*) and "Krogan" (*16*). For literature-curated interactions, we only considered those curated from two or more publications ("LC-multiple") (*20*), which we considered of higher quality than those curated from a single publication.

To experimentally compare the quality of these datasets, we selected a representative sample of ~200 protein interaction pairs from each one and tested them by means of two independent interaction assays, Y2H and a yellow fluorescent protein complementation assay (PCA) (*21*) [Supporting Online Material (SOM) I]. In PCA, bait and prey proteins are fused to non-fluorescent fragments of yellow fluorescent protein that, when brought in close proximity by interacting proteins, reconstitute a fluorescent protein in mammalian cells. In contrast, reconstitution of a transcription factor in Y2H experiments takes place in the nucleus of yeast cells. In terms of assay designs, Y2H and PCA can be considered as orthogonal assays and can be used to validate each other's results.

No single assay is expected to detect 100% of genuine interactions, and the actual fraction of positives detected is inherently linked to the stringency at which the assay is implemented. To identify the optimal scoring condition of each assay we selected a set of ~100 well-documented yeast protein-protein interaction pairs ["positive reference set" (PRS)] and a set of ~100 random pairs ["random reference set" (RRS)] (Fig. 1B; SOM II). Because RRS pairs were picked uniformly from the $14x10^6$ possible pairings of proteins within our yeast ORFeome collection (*22*) (excluding those reported as interacting), these pairs are extremely unlikely to be interacting.

Sampled pairs from binary Uetz-screen and Ito-core datasets tested positive at levels as high as the positive control PRS, demonstrating their high quality (Fig. 1C). A sample of literature-curated LC-multiple interactions tested slightly lower with Y2H, while being indistinguishable by PCA (Fig. 1C), demonstrating that high-throughput Y2H datasets can be comparable in quality to literature-curated information. In striking contrast, sampled pairs from Ito-noncore tested at levels similar to the negative control RRS, confirming the low quality of this particular dataset (Fig. 1C).

Sampled pairs from Gavin and Krogan high-throughput AP/MS datasets tested poorly in our two binary interaction assays (Fig. 1C), albeit at levels similar to Munich Information Center for Protein Sequences (MIPS) complexes, a widely-used gold standard (*23*). This observation demonstrates that, at least for detecting binary interactions, Y2H performs better than AP/MS

Our experimental data quality assessment shows that binary Uetz-screen, Ito-core, and LC-multiple datasets are of high quality, while Ito-noncore should not be used. AP/MS datasets, although of intrinsically good quality (*15, 16*), should be used with caution when binary interaction information is needed.

Our experimental results contrast strikingly with computational analyses that suggested that high-throughput Y2H datasets contain more false positives than literature-curated or high-throughput AP/MS datasets (*24, 25*). In computational analyses, the quality of a dataset is often determined by the fraction of interactions also present in a pre-defined gold standard set (*24*). Generally, MIPS complexes have been considered as gold standard with all proteins constituting a given complex modeled as interacting with each other. Such modeling results in limited and biased sampling issues against binary interactions since not all proteins in a complex contact each other directly (fig. S1), and not all direct physical interactions occur within complexes (fig. S2; SOM III). Hence, while MIPS complexes are appropriate for benchmarking co-complex membership datasets, they are not for binary interaction datasets. This distinction is corroborated by the poor experimental confirmation rate of pairs from MIPS complexes using binary assays (Fig. 1C).

To computationally re-examine the quality of existing yeast interactome datasets we assembled a binary gold standard set ("Binary-GS") of 1,318 high-confidence physical binary interactions (Fig. 1B; SOM III). Binary-GS includes direct physical interactions within well-established complexes as well as conditional interactions (*e.g.*, dependent on posttranslational modifications) and thus represents well-documented direct physical interactions in the yeast interactome (*26*). When measured against Binary-GS, the quality of high-throughput Y2H datasets (with the exception of Ito-noncore) was substantially better (SOM IV and V) than that of high-throughput AP/MS datasets (Fig. 1D). Our results demonstrate the distinct nature of binary and co-complex data. Generally, Y2H datasets contain high quality direct binary interactions, whereas AP/MS co-complex datasets are

composed of direct interactions mixed with preponderant indirect associations (SOM VI).

The proteome-wide binary datasets, Uetz-screen and Ito-core, contain 682 and 843 interactions, respectively (*2, 3*). The overlap between these two datasets appears low (*3, 24*): 19% of Uetz-screen and 15% of Ito-core interactions were detected in the other dataset. Given our demonstration of high quality for these datasets (Fig. 1C, D), we conclude that the small overlap stems primarily from low sensitivity (*i.e.*, many false negatives) rather than from low specificity (*i.e.*, many false positives as previously suggested).

Several factors might affect sensitivity. First, the space of pair-wise protein combinations actually tested in each dataset might have been considerably different. We refer to the fraction of all possible pairs tested in a given screen as the "completeness". For example, missing 10% of ORFs in each mapping project could reduce the common tested space down to 66% [(0.9x0.9) x (0.9x0.9)] of all possible pair-wise combinations. Second, different protein interaction assays or even different versions of the same assay detect different subsets of pairs out of all possible interactions, explaining partly the limited overlap between datasets obtained with different Y2H versions. For any assay, the "assay-sensitivity" is estimated as the fraction of PRS interactions detected, which for our Y2H assay was determined empirically to be ~20% (Fig. 1C). Finally, when screening tens if not hundreds of millions of protein pairs in any tested space, that search space might need to be sampled multiple times to report all or nearly all interactions detectable by the assay used. The fraction of all theoretically detectable interactions by a particular assay found in a given experiment is its "sampling-sensitivity". These three parameters fully account for the seemingly small overlap between Ito-core and Uetz-screen (SOM VII), demonstrating that a large fraction of the *S. cerevisiae* binary interactome remains to be mapped. Therefore, we carried out a new proteome-scale yeast high-throughput Y2H screen (fig. S3).

We used 5,796 Gateway-cloned ORFs available in the yeast MORF collection (*22*). After subcloning these ORFs into Y2H vectors and removing auto-activators (*27, 28*), our search space became 3,917 DB-Xs against 5,246 AD-Ys, representing a completeness of 77% (Fig. 2A; SOM VI), comparable to that of recent AP/MS datasets (*15*) (~78%; SOM VI).

To address sampling-sensitivity, we determined what fraction of all detectable interactions is found in each pass after eight trials in a search space of 658 DB-X and 1,249 AD-Y ORFs. A single trial identified about 60% of all possible interactions that can be detected with our high-throughput Y2H, whereas three to five repeats were required to obtain 80-90% (Fig. 2B; SOM VI). We decided to screen the whole search space three times independently to yield an estimated sampling-sensitivity of 85% (Fig. 2B). In total ~88,000 colonies were picked, of which 21,432 scored positive upon more detailed phenotyping (SOM I). After identifying all putative interaction pairs by sequencing, phenotypically retesting them using fresh cultures from archival stocks, and eliminating *de novo* auto-activators (*28*), we obtained a final dataset, "CCSB-YI1", of 1,809 interactions among 1,278 proteins.

To validate the overall quality of CCSB-YI1, we tested 94 randomly-chosen interactions by PCA and mammalian protein-protein interaction trap (MAPPIT; SOM I) (*21, 29*). MAPPIT takes place at the mammalian cell membrane and measures interactions via activation of STAT3-dependent reporter expression. Using both PCA and MAPPIT the confirmation rate of CCSB-YI1 was similar to those of Ito-core and Uetz-screen (Fig. 1C). The precision [*i.e.*, fraction of true positives in the dataset (*30*)] of CCSB-YI1 is estimated at 94-100% (Fig. 2C; fig. S4; SOM VI). Additionally, the performance of our high-throughput Y2H approach was confirmed via a larger RRS of 1,000 random pairs (*30*) (Fig. 1B), none of which tested positive (SOM II).

The overlaps of Uetz-screen (27%) and Ito-core (35%) with CCSB-YI1 (Fig. 2D) can be explained by the completeness, assay- and sampling-sensitivity of the three experiments (SOM VII) and agree well with the results of the pair-wise confirmation of those two datasets (Fig. 1C). Similar principles apply to other large-scale experiments such as AP/MS, likely accounting for the low overlap between Krogan and Gavin (~25%; fig. S5B).

Factoring in completeness, precision, assay-, and sampling-sensitivity, we estimated that the yeast binary interactome consists of ~18,000±4,500 interactions (SOM VI), experimentally validating previous computational estimates of 17,000 to 25,000 interactions (*31, 32*). To obtain a more comprehensive map of the binary yeast interactome we combined the three available high-quality proteome-scale Y2H datasets (SOM VII). The union of Uetz-screen, Ito-core, and CCSB-YI1, "Y2H-union", contains 2,930 binary interactions among 2,018 proteins, which, according to our empirical estimate of the interactome size, represents ~20% of the whole yeast binary interactome (Fig 3A).

We re-examined global topological features of this new yeast interactome network, facing lower risk of over-interpreting properties due to limited sampling and various biases in the data (*18*). To contrast topological properties of the binary Y2H-union network with that of the co-complex network, we used an integrated AP/MS dataset (*33*), which was generated by combining raw high-throughput AP/MS data (*15, 16*). This "Combined-AP/MS" dataset, composed of 9,070 co-complex membership associations between 1,622 proteins, attempts to model binary interactions from co-complex data (Fig. 3A).

As found previously for other macromolecular networks, the connectivity or "degree" distribution of all three datasets is best approximated by a power-law (*34*) (fig. S6; SOM VIII). Highly connected proteins, or "hubs", are reportedly more likely encoded by essential genes than less connected proteins (*35*). Surprisingly, Y2H-union lacked any correlation between degree and essentiality (Fig. 3B). This discrepancy might stem from biases in the datasets available at the time of the original observation: interactions reported in Uetz *et al.* (Uetz-array and Uetz-screen) and literature-curated interactions. Although Uetz-array is of high quality (fig. S7), its experimental design could negatively influence network analyses. Most hub proteins in Uetz-array were found as baits (fig. S8) and the percentage of essential proteins in the 193 bait proteins is two times higher (34.7%) than that of all protein-encoding ORFs in the yeast genome (18.4%), explaining the high correlation between degree and essentiality (Fig. 3C). Likewise, literature-curated interactions seem prone to sociological and other inspection biases (SOM VII). Thus, we refrain heretofore from using LC-multiple in our further topological and biological analyses. No significant correlation between degree of connectedness and essentiality was observed in any of the three proteome-wide high-throughput binary datasets available today (*i.e.*, Ito-core, Uetz-screen, and CCSB-YI1; Fig. 3C), as well as new versions of our *C. elegans* and human interactome maps (fig. S9; SOM IX).

Hub proteins instead relate to pleiotropy, the number of phenotypes observed as a consequence of gene knock-out (SOM I). There was a significant correlation in Y2H-union between connectivity and the number of phenotypes observed in global phenotypic profiling analyses of yeast genes (*36*) (Fig. 3D). Thus the number of binary physical interactions mediated by a protein seems to better correlate with the number of cellular processes in which it participates than its essentiality. The correlation between degree and number of phenotypes is not observed in Combined-AP/MS, likely because co-complex associations reflect the size of protein complexes more than the number of processes they might be involved in.

We confirmed the concept of modularity in the yeast interactome network, whereby date hubs that dynamically interact with their partners appear particularly central to global connectivity while static party hubs appear to function locally in specific biological modules (*37*). The proportion of date and party hubs is strikingly different between Y2H-union and Combined-AP/MS (Fig. 3E). There are significantly more date hubs in the binary network, whereas party hubs are prevalent in the co-complex network. In the binary network, date hubs are crucial to the topological integrity of the network, while party hubs have minimal effects. However, in the co-complex network, date and party hubs affect the topological integrity of the network equally, likely because most hubs in Combined-AP/MS reside in large stable complexes, while hubs in Y2H-union preferentially connect diverse cellular processes.

Surprisingly, essential proteins strongly tended to interact with each other (Fig. 4A; SOM IX). Concentrating on the subnetwork formed by interactions mediated by and among essential proteins (fig. S10), we found a giant component whose size is much larger than expected by chance (Fig. 4B). To better understand the clustering of essential proteins, we examined the interacting essential protein pairs that are also reported to be in the same complex, finding 106 interacting essential protein pairs, a greater number than expected by chance (Fig. 4C; SOM IX).

We investigated the overall relationships between Y2H-union and Gene Ontology (GO) attributes (*38*), phenotypic and expression profiling similarities (*39*), and transcriptional regulatory networks (*40*). Both Y2H-union and Combined-AP/MS show significant enrichment (all $P<10^{-10}$) for functionally similar pairs in all three GO branches (Fig. 5A) (*41*). There is also significant enrichment of positive correlations of phenotypic profiles (*36*) between interacting pairs in both datasets (Fig. 5B; fig. S11). Such interactions supported by strong phenotypic information constitute likely possibilities of functional relationships. Lastly, both datasets are significantly enriched with pairs co-expressed across many conditions (fig. S12), although Combined-AP/MS shows higher enrichment (Fig. 5C), agreeing well with the different nature of the two assays: AP/MS aims at detecting stable complexes whereas Y2H tends to detect more transient and condition specific protein interactions. This observation is further supported by enrichment of kinase-substrate pairs in Y2H-union (SOM X; fig. S13).

To explore the mechanisms behind co-expression of interacting protein pairs we combined transcriptional regulatory networks with interactome network information (*40*). Interacting proteins in both networks showed a tendency to be co-regulated by common transcription factors (TFs; Fig. 5D). Similarly to what we observed in the co-expression correlation analysis (Fig. 5C), the enrichment for interacting pairs in Combined-AP/MS was significantly higher than that of Y2H-union. Strikingly, we observed a significant enrichment of protein-protein interactions between TFs involved in a common "multi-input motif" (*42, 43*) (MIM, where multiple TFs co-regulate a given set of genes; Fig. 5D; SOM X). The fraction of co-regulating TF pairs is much higher in the binary interactome than in the co-complex network, suggesting that various TFs function together to form transient complexes to differentially regulate transcriptional targets (*44*).

These observations suggest that our binary interactome dataset is enriched in transient or condition-specific

interactions linking different subcellular processes and molecular machines. To further explore this possibility we calculated "edge-betweenness" for each interaction in a merged network of all available interactions (SOM XI), measuring the number of shortest paths between all protein pairs that traverse a given edge. The higher edge-betweenness of interactions from Y2H-union shows the tendency of Y2H to detect key interactions outside of complexes, significantly more often than AP/MS (Fig. 5D). Several examples of such complex-to-complex connectivity are evident in a complete map of MIPS complexes connected by Y2H interactions (fig. S14).

Overall, we infer that Y2H interrogates a different subspace within the whole interactome than AP/MS, and Y2H interactions represent key connections between different complexes and pathways. Y2H and AP/MS provide orthogonal information about the interactome and are both vital to obtain a complete picture of cellular protein-protein interaction networks.

References and Notes

1. M. Fromont-Racine, J. C. Rain, P. Legrain, *Nat. Genet.* **16**, 277 (1997).
2. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
3. T. Ito *et al.*, *Proc. Natl. Acad. Sci. USA* **98**, 4569 (2001).
4. A. J. Walhout *et al.*, *Science* **287**, 116 (2000).
5. J. Reboul *et al.*, *Nat. Genet.* **34**, 35 (2003).
6. S. Li *et al.*, *Science* **303**, 540 (2004).
7. L. Giot *et al.*, *Science* **302**, 1727 (2003).
8. F. Colland *et al.*, *Genome Res.* **14**, 1324 (2004).
9. J. F. Rual *et al.*, *Nature* **437**, 1173 (2005).
10. U. Stelzl *et al.*, *Cell* **122**, 957 (2005).
11. G. Butland *et al.*, *Nature* **433**, 531 (2005).
12. M. Arifuzzaman *et al.*, *Genome Res.* **16**, 686 (2006).
13. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
14. Y. Ho *et al.*, *Nature* **415**, 180 (2002).
15. A. C. Gavin *et al.*, *Nature* **440**, 631 (2006).
16. N. J. Krogan *et al.*, *Nature* **440**, 637 (2006).
17. R. M. Ewing *et al.*, *Mol. Syst. Biol.* **3**, 89 (2007).
18. J. D. Han, D. Dupuy, N. Bertin, M. E. Cusick, M. Vidal, *Nat. Biotechnol.* **23**, 839 (2005).
19. D. Scholtens, M. Vidal, R. Gentleman, *Bioinformatics* **21**, 3548 (2005).
20. T. Reguly *et al.*, *J. Biol.* **5**, 11 (2006).
21. I. Remy, S. W. Michnick, *Methods Mol. Biol.* **261**, 411 (2004).
22. D. M. Gelperin *et al.*, *Genes Dev.* **19**, 2816 (2005).
23. H. W. Mewes *et al.*, *Nucleic Acids Res.* **34**, D169 (2006).
24. C. von Mering *et al.*, *Nature* **417**, 399 (2002).
25. J. S. Bader, A. Chaudhuri, J. M. Rothberg, J. Chant, *Nat. Biotechnol.* **22**, 78 (2004).
26. H. Yu *et al.*, *Genome Res.* **14**, 1107 (2004).
27. P. O. Vidalain, M. Boxem, H. Ge, S. Li, M. Vidal, *Methods* **32**, 363 (2004).
28. A. J. Walhout, M. Vidal, *Genome Res.* **9**, 1128 (1999).
29. S. Eyckerman *et al.*, *Nat. Cell Biol.* **3**, 1114 (2001).
30. R. Jansen, M. Gerstein, *Curr. Opin. Microbiol.* **7**, 535 (2004).
31. A. Grigoriev, *Nucleic Acids Res.* **31**, 4157 (2003).
32. R. Jansen *et al.*, *Science* **302**, 449 (2003).
33. S. R. Collins *et al.*, *Mol. Cell. Proteomics* **6**, 439 (2007).
34. A. L. Barabási, R. Albert, *Science* **286**, 509 (1999).
35. H. Jeong, S. P. Mason, A. L. Barabasi, Z. N. Oltvai, *Nature* **411**, 41 (2001).
36. A. M. Dudley, D. M. Janse, A. Tanay, R. Shamir, G. M. Church, *Mol. Syst. Biol.* **1**, 0001 (2005).
37. J. D. Han *et al.*, *Nature* **430**, 88 (2004).
38. Gene Ontology Consortium, *Nucleic Acids Res.* **36**, D440 (2008).
39. M. Vidal, *Cell* **104**, 333 (2001).
40. H. Yu, M. Gerstein, *Proc. Natl. Acad. Sci. USA* **103**, 14724 (2006).
41. H. Yu, R. Jansen, G. Stolovitzky, M. Gerstein, *Bioinformatics* **23**, 2163 (2007).
42. T. I. Lee *et al.*, *Science* **298**, 799 (2002).
43. R. Milo *et al.*, *Science* **298**, 824 (2002).
44. N. M. Luscombe *et al.*, *Nature* **431**, 308 (2004).
45. H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, *PLoS Comput. Biol.* **3**, e59 (2007).

46. Supported by funds from the W.M. Keck Foundation awarded to M.V. and F.P.R.; by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative awarded to M.V. and CCSB; by National Institutes of Health (NIH) grant (R01-HG001715) awarded to M.V. and F.P.R.; by National Institutes of Health (NIH) grants (U01-A1070499-01 and U56-CA113004) awarded to A.-L.B.; by grants from the University of Ghent (GOA12051401) and the Fund for Scientific Research Flanders (FWO-V G.0031.06) awarded to J.T.; by a grant from the National Cancer Institute of Canada awarded to C.B.; and by an NIH grant (HG003224) awarded to F.P.R. I.L. is a postdoctoral fellow with the FWO-V. M.V. is a "Chercheur Qualifié Honoraire" from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium). We thank members of our laboratories for helpful discussions, and Agencourt Biosciences for sequencing assistance. All datasets can be downloaded from our website: http://interactome.dfci.harvard.edu/S_cerevisiae.

Supporting Online Material

SOM Text
Figs. S1 to S35
Tables S1 to S5
References

4 April 2008; accepted 1 August 2008
Published online 21 August 2008; 10.1126/science.1158684
Include this information when citing this paper.

Fig. 1. Evaluation of *S. cerevisiae* protein-protein interaction datasets. (**A**) Number of interactions reported in various large-scale *S. cerevisiae* protein-protein interaction datasets. (**B**) Schema of pipeline used to assemble binary positive and random reference sets. (**C**) Fraction of a random sample of interactions from each dataset confirmed by Y2H and PCA. (**D**) Fraction of positives in each dataset calculated using MIPS and Binary-GS. (Error bars indicate standard error).

Fig. 2. Large-scale Y2H interactome screen. (**A**) Completeness of the Y2H screen. (**B**) Sampling-sensitivity of CCSB Y2H screens measured by screening a subspace multiple times. (**C**) Fraction of protein pairs in PRS, RRS, and CCSB-YI1 that test positive by PCA, MAPPIT or Y2H. (**D**) Overlaps between three high-quality large-scale *S. cerevisiae* Y2H datasets ($*P<10^{-7}$).

Fig. 3. Network analysis of Y2H-union, Combined-AP/MS and LC-multiple datasets. (**A**) Network representations. Shown are relationships between increasing degree of a gene product and (**B**) the fraction of essential genes with the corresponding degree, (**C**) the fraction of essential genes with the corresponding degree for Y2H datasets, (**D**) the number of phenotypes associated with deletion of the encoding gene. (**E**) Contribution of date hubs and party hubs as measured by change in the characteristic path length after simulated removal of edges by deleting the indicated types of nodes. Inset: fraction of date hubs and party hubs for each dataset.

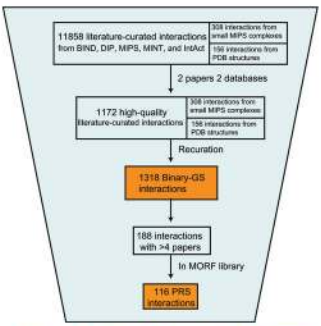
Fig. 4. Clustering of essential proteins. (**A**) Average fraction of essential proteins among proteins whose distance are equal to *d* from a protein selected from essential, non-essential and all proteins. (**B**) Giant component size of network formed by essential proteins (arrow) compared to 100,000 random networks of same topological properties. (**C**) The number of interacting essential proteins that are also found in the same complex compared to 10,000 random selections of proteins of the same number as essential proteins (SOM IX).

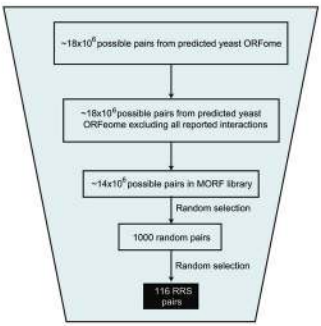
Fig. 5. Biological features of yeast interactome datasets. (**A**) Enrichment of interacting protein pairs (relative to random) that share GO annotations in the biological process, cellular component and molecular function branches of GO ontology. (**B**) Pearson correlation coefficient (*PCC*) of phenotypic profiles between interacting pairs in different datasets. (**C**) Co-expression correlation between interacting pairs. (**D**) Left panel: enrichment of interacting proteins as targets of a common TF (co-regulated), and enrichment of interacting TFs in a common MIM (co-regulating) ($*P<10^{-3}$). Right panel: fraction of bottlenecks from each dataset in the combined network (SOM XI). Top 10% of edges with the highest betweenness are defined as "bottlenecks" (*45*).

A

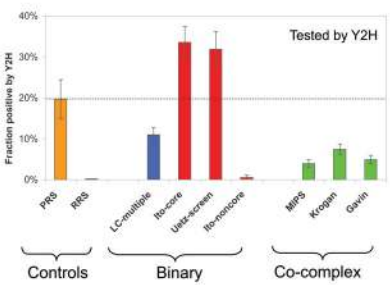
Reference	Uetz *et al.* (2)		Ito *et al.* (3)		Gavin *et al.* (15)	Krogan *et al.* (16)	Reguly *et al.* (20)
Dataset name	Uetz-screen	Uetz-array	Ito-core	Ito-noncore	Gavin	Krogan	LC-multiple
Number of reported interactions	682	281	843	3632	6942	7123	2925
	Binary				Co-complex		Literature

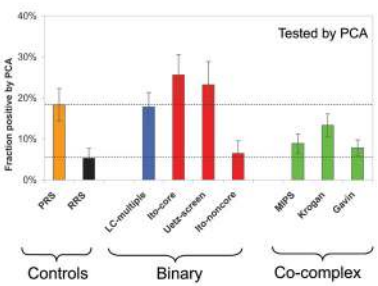
B





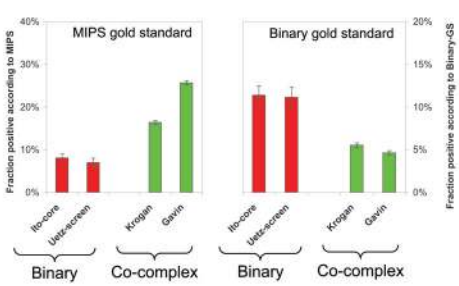
C



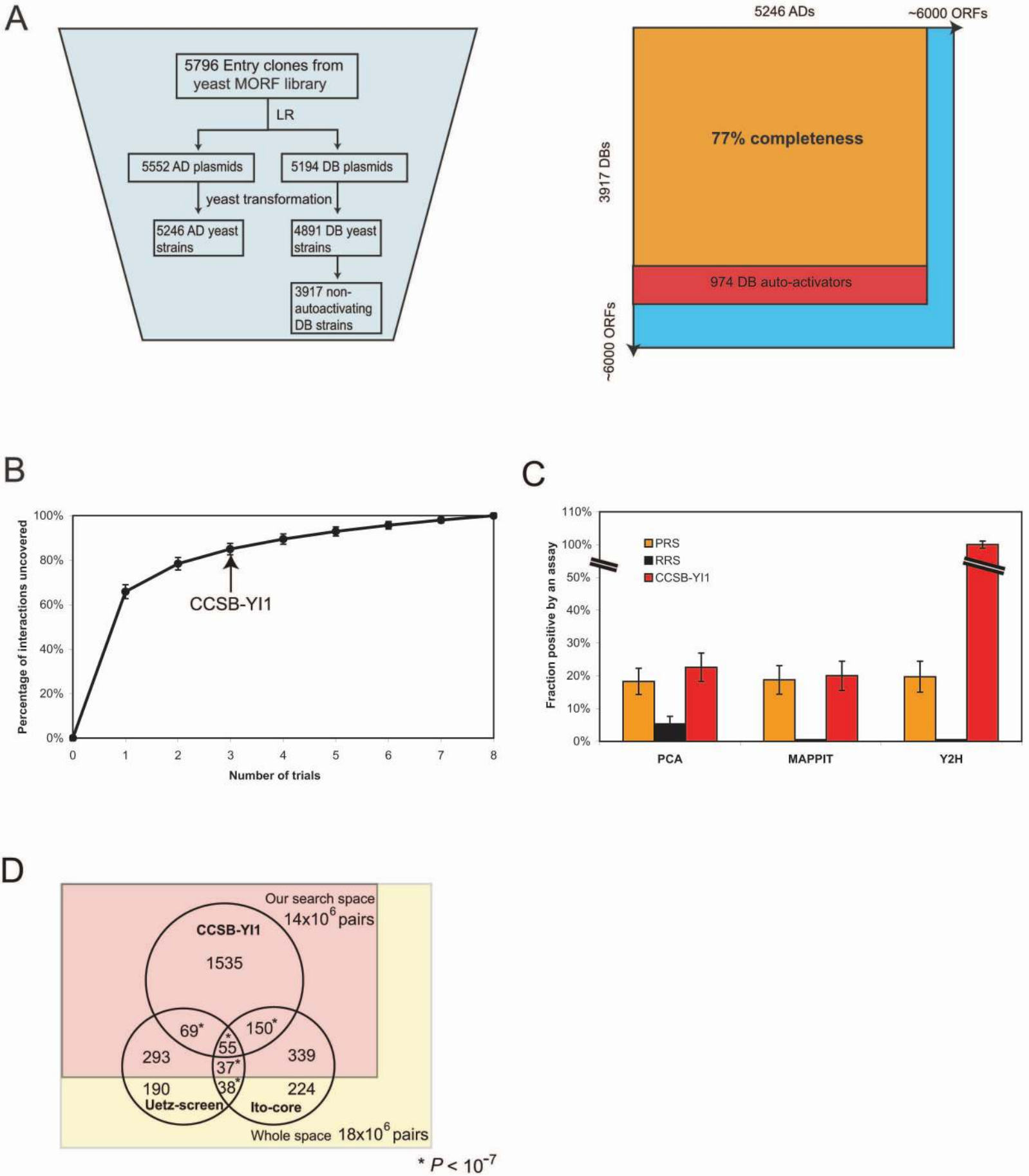


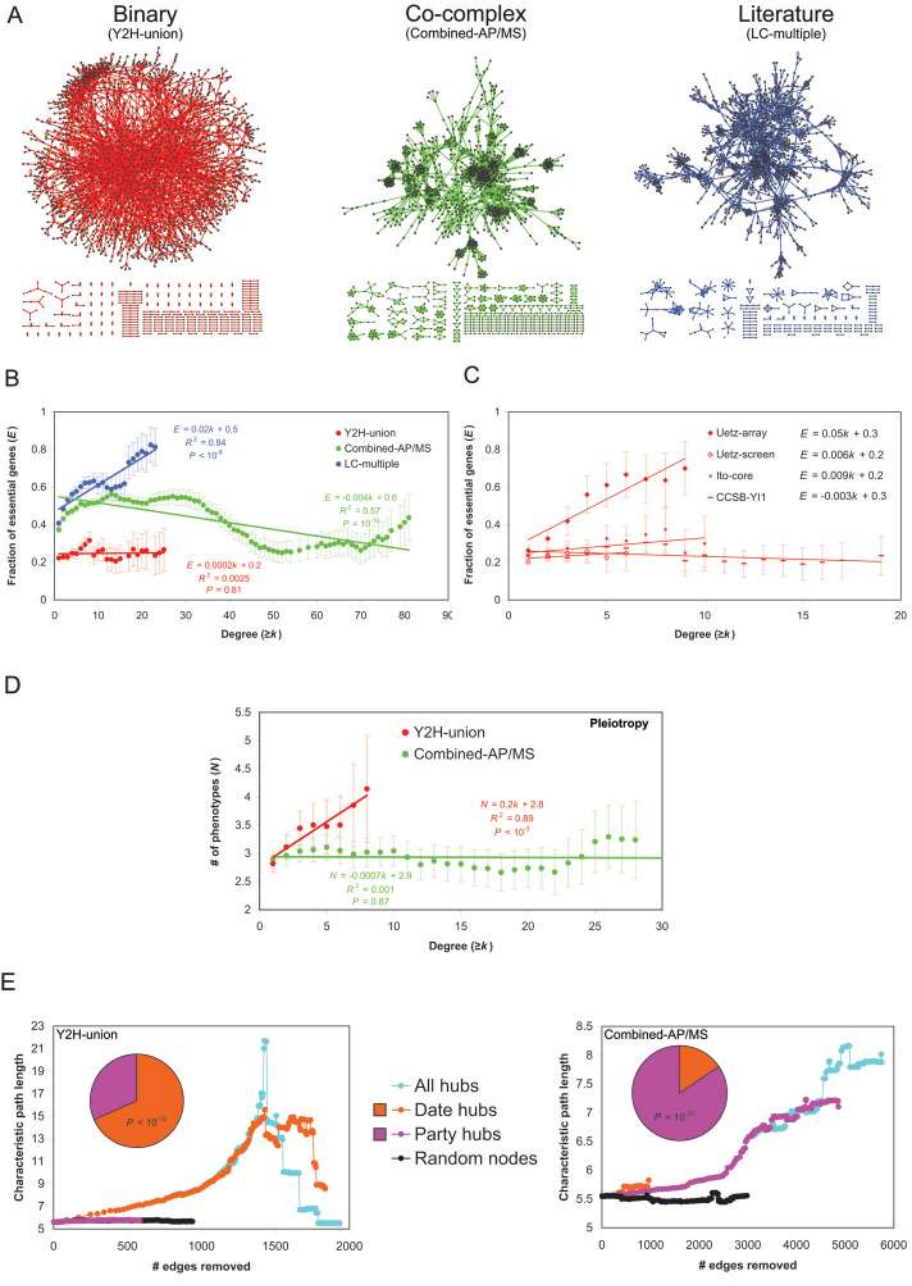
Brief summary (For detecting binary interactions)	Representative datasets	*P*-values Y2H	*P*-values PCA
1. Y2H is comparable to LC-multiple	Uetz-screen ≥ LC-multiple	$P<10^{-16}$	$P=0.1$
2. Ito-noncore is of extremely low quality	Ito-noncore ≈ RRS	$P=0.9$	$P=0.2$
3. Y2H performs better than AP/MS	Uetz-screen > Krogan	$P<10^{-30}$	$P<10^{-3}$
4. Y2H performs better than MIPS	Uetz-screen > MIPS	$P<10^{-41}$	$P<10^{-5}$

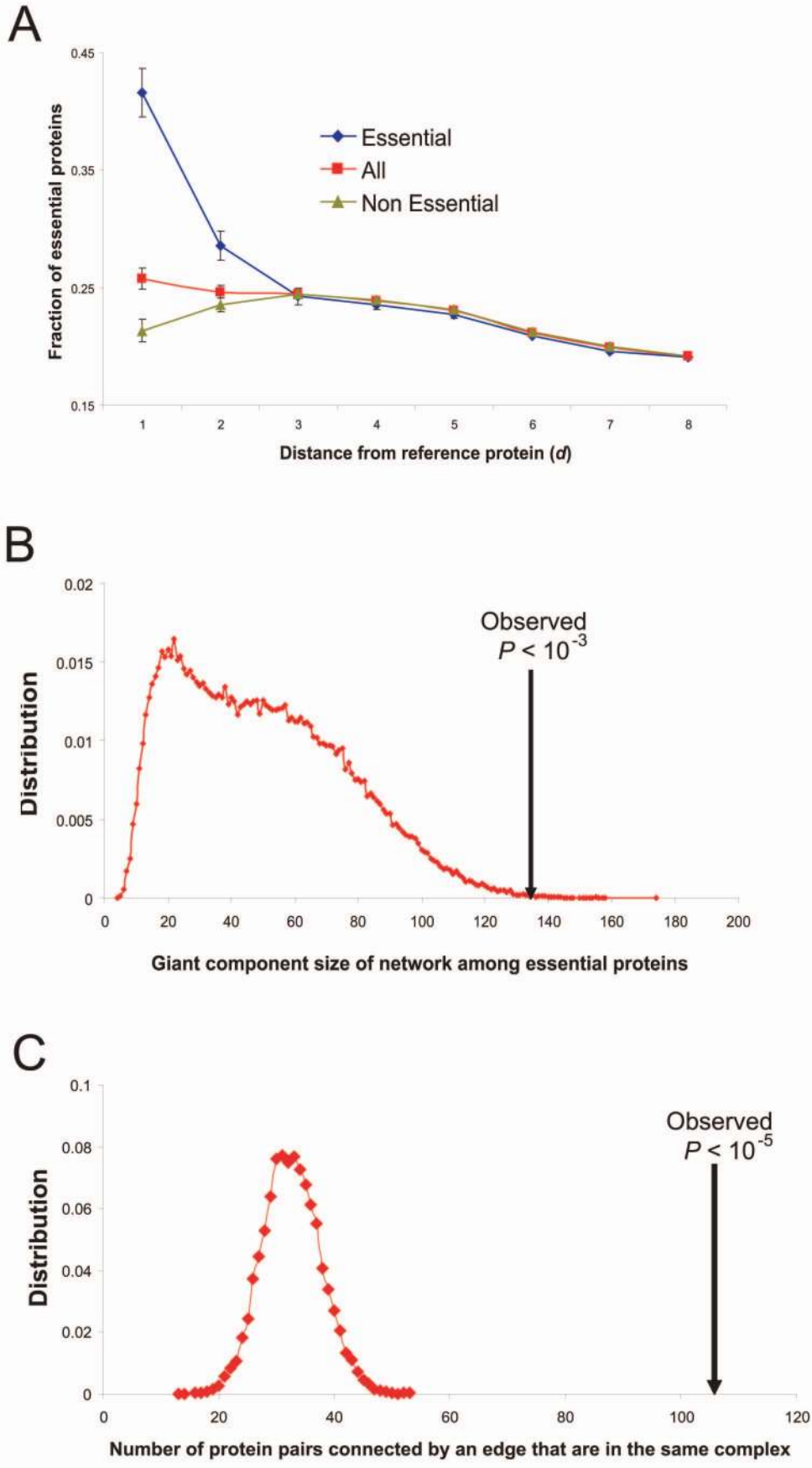
D

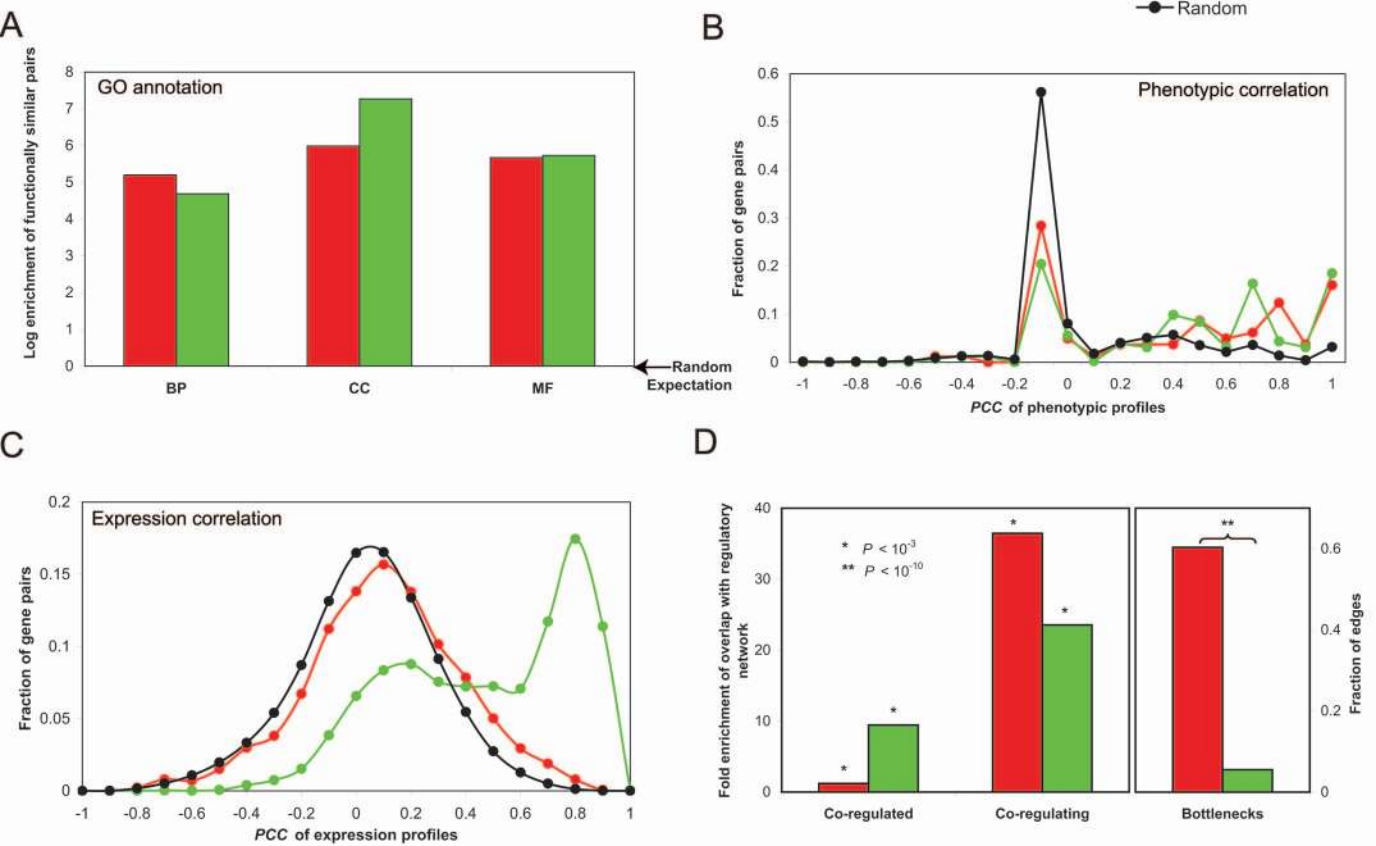


Brief summary	Representative datasets	*P*-values
1. AP/MS performs well at detecting co-complex associations according to MIPS	Uetz-screen < Krogan	$P<10^{-12}$
2. Y2H performs well at detecting binary interactions according to Binary-GS	Uetz-screen > Krogan	$P<10^{-5}$


A
5796 Entry clones from yeast MORF library
LR
5552 AD plasmids
5194 DB plasmids
yeast transformation
5246 AD yeast strains
4891 DB yeast strains
3917 non-autoactivating DB strains
5246 ADs
~6000 ORFs
77% completeness
3917 DBs
974 DB auto-activators
~6000 ORFs
B
Percentage of interactions uncovered
100%
80%
60%
40%
20%
0%
0
1
2
3
4
5
6
7
8
Number of trials
CCSB-YI1
C
Fraction positive by an assay
110%
100%
50%
40%
30%
20%
10%
0%
PRS
RRS
CCSB-YI1
PCA
MAPPIT
Y2H
D
Our search space
14x10^6 pairs
CCSB-YI1
1535
69*
150*
55*
293
339
37*
38*
190
224
Uetz-screen
Ito-core
Whole space 18x10^6 pairs
* P < 10^-7


A
Binary
(Y2H-union)
Co-complex
(Combined-AP/MS)
Literature
(LC-multiple)
B
Fraction of essential genes (E)
Degree (≥k)
Y2H-union
Combined-AP/MS
LC-multiple
E = 0.02k + 0.5
R² = 0.84
P < 10⁻⁹
E = -0.004k + 0.6
R² = 0.57
P < 10⁻¹⁵
E = 0.0002k + 0.2
R² = 0.0025
P = 0.81
C
Fraction of essential genes (E)
Degree (≥k)
Uetz-array
Uetz-screen
Ito-core
CCSB-YI1
E = 0.05k + 0.3
E = 0.006k + 0.2
E = 0.009k + 0.2
E = -0.003k + 0.3
D
# of phenotypes (N)
Degree (≥k)
Pleiotropy
Y2H-union
Combined-AP/MS
N = 0.2k + 2.8
R² = 0.89
P < 10⁻⁵
N = -0.0007k + 2.9
R² = 0.001
P = 0.87
E
Y2H-union
Characteristic path length
edges removed
P < 10⁻¹⁶
All hubs
Date hubs
Party hubs
Random nodes
Combined-AP/MS
Characteristic path length
edges removed
P < 10⁻¹⁶


A
0.45
0.35
0.25
0.15
Fraction of essential proteins
Essential
All
Non Essential
1
2
3
4
5
6
7
8
Distance from reference protein (d)
B
0.02
0.015
0.01
0.005
0
Distribution
Observed
P < 10-3
0
20
40
60
80
100
120
140
160
180
200
Giant component size of network among essential proteins
C
0.1
0.08
0.06
0.04
0.02
0
Distribution
Observed
P < 10-5
0
20
40
60
80
100
120
Number of protein pairs connected by an edge that are in the same complex


A
GO annotation
Log enrichment of functionally similar pairs
BP
CC
MF
Random Expectation
B
Random
Phenotypic correlation
Fraction of gene pairs
PCC of phenotypic profiles
C
Expression correlation
Fraction of gene pairs
PCC of expression profiles
D
* P < 10-3
** P < 10-10
Fold enrichment of overlap with regulatory network
Fraction of edges
Co-regulated
Co-regulating
Bottlenecks