www.sciencemag.org/cgi/content/full/1153878/DC1

Supporting Online Material for

An in Vivo Map of the Yeast Protein Interactome

Kirill Tarassov, Vincent Messier, Christian R. Landry, Stevo Radinovic,
Mercedes M. Serna Molina, Igor Shames, Yelena Malitskaya, Jackie Vogel,
Howard Bussey, Stephen W. Michnick*

*To whom correspondence should be addressed. E-mail: stephen.michnick@umontreal.ca

Published 8 May 2008 on *Science* Express
DOI: 10.1126/science.1153878

This PDF file includes:

Materials and Methods
Figs. S1 to S11
References

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/1153878/DC1)

Tables S1 to S9 as zipped archives
Data Set S1 as zipped archive
Software package for viewing Data Set S1; versions for Windows, Mac OSX, and Linux
PDF documentation for above software package

SUPPLEMENTARY ONLINE MATERIAL

AN *IN VIVO* MAP OF THE YEAST PROTEIN INTERACTOME

Kirill Tarassov, Vincent Messier, Christian R. Landry, Stevo Radinovic, Mercedes M. Serna Molina, Igor Shames, Yelena Malitskaya, Jackie Vogel, Howard Bussey and Stephen W. Michnick

Département de Biochimie
Université de Montréal
C.P. 6128, Succursale centre-ville
Montréal, Québec H3C 3J7
Canada

Material and Methods and Supplementary Information

1. Adaptation of the mDHFR Protein-fragment Complementation Assay for studies in yeast. The mDHFR PCA (henceforth DHFR PCA) was previously developed for *E. coli*, plant protoplasts and mammalian cell lines (*1-3*). To adapt the DHFR PCA for high-throughput screening in *Saccharomyces cerevisiae* we created a double mutant (L22F and F31S) that is 10,000 times less sensitive to methotrexate than wildtype scDHFR, while retaining full catalytic activity (*4*). The L22F-F31S double mutant was created by introducing by site-directed mutagenesis the L22F mutation into the existing F31S mutant DHFR PCA N-terminal fragment (F[1,2]) previously developed for studies in mammalian cells (*3*). The mutant was designed so that methotrexate inhibition of the wildtype scDHFR activity would be complemented by the activity of the reconstituted DHFR PCA reporter.

In order for a PCA to minimally perturb the natural kinetics of protein-protein interactions and to prevent trapping of non-specific complexes, it has to be reversible; the reporter protein has to unfold and the fragments dissociate upon the disruption of a protein complex. PCAs based on the *Gaussia princeps* and *Renilla reniformis* luciferase have been shown to have this property whereas those based on GFP variants are irreversible (*4, 5*). To directly assay the reversibility of the DHFR PCA, we used an *in vitro* assay that uses the cAMP dependent dissociation of regulatory (Bcy1) and catalytic (Tpk2) subunits of the yeast homologue of serine/threonine kinase PKA. In this assay, binding of the Bcy1 regulatory subunit to resin-immobilized cAMP is probed and whether the Tpk2 catalytic subunit is dissociated or remains associated with Bcy1 can be assessed (*5, 6*). These experiments allow us to directly distinguish two possible outcomes of Bcy1:Tpk2 dissociation (Fig. S10 A): i), the PCA is a trap; that is, while a PKA complex may dissociate on binding to the cAMP resin conjugate, the PCA fragments do not unfold and thus both catalytic and regulatory proteins remain bound to the resin; ii), PKA subunits dissociate and the PCA fragments unfold and separate, resulting in the release of the catalytic domain from the cAMP resin. In this case, the regulatory subunit remains bound to the resin while the catalytic subunit would be found in supernatant. We tested the reversibility of the DHFR PCA in direct comparison to the Rluc and “Venus” YFP PCAs, which we previously showed to be respectively reversible and non-reversible (*5*). We constructed two sets of yeast expression vectors harboring fusions of Bcy1 and Tpk2 to complementary fragments of the Rluc, YFP and DHFR PCAs under control of a constitutive TEF promoter. Bcy1 and Tpk2 genes were PCR amplified from the *S. cerevisiae* genome and subcloned into yeast expression vectors harboring mDHFR, Renilla luciferase (Rluc) or “Venus” YFP PCA fragments fused 3’ to Zip-linker, replacing the Zip sequences (p413-TEF-Zip-linker-F[3], p415-TEF-Zip-linkers-F[1,2], p413-TEF-Zip-linker-hRluc-F[1] and p415-TEF-Zip-linker-hRlucF[2], p413-TEF-Zip-linker-Venus YFP F[1] and p415-TEF-Zip linker-Venus YFP F[2]). Approximately 100 ng of each of the complementary PCA fragments expression vectors were cotransformed into *S. cerevisiae* BY4743 strain (*MAT*a/α *his3*Δ*/his3*Δ *leu2*Δ*/leu2*Δ *lys2*Δ*/LYS2 MET15/met15*Δ *ura3*Δ*/ura3*Δ) and positive clones were selected in synthetic complete medium (SC) –methionine, -lysine, -uracil and –histidine for p413/p416 transformed strains. BY4743 diploid strains harboring plasmids encoding the three

DHFR, Rluc or “Venus” YFP Bcy1-Tpk2 PCAs were grown in 5 ml of synthetic complete medium (SC) (-met, -lys, -his, -leu) to an OD_{600} of 1.0, harvested and treated with 200 units of lyticase (Sigma-Aldrich, St-Louis, MO) for 2 hrs at 30°C to digest the cell wall. We extracted soluble lysate and incubated with the cAMP resin followed by a series of washes and then probed supernatant and bound fractions with antibodies that can bind to each PCA fragment (Fig. S10 B). Specifically, the protoplasm was harvested at 500xg for 30 minutes at 4°C and resuspended in 0.1% volume of the original culture volume of phosphate-buffered saline (PBS) (pH 7.4) + 1% (v/v) Triton-X-100 + protease inhibitors (100μg/ml) PMSF and MiniComplet EDTA free protease inhibitor tablets (Roche, Indianapolis, IN) and stored on ice for 30 minutes. The lysate was clarified by centrifuging at 12,000xg for 30 minutes at 4°C. The soluble fraction of each lysate was separated into two equal aliquots (volume of 250 μL). The first aliquot was incubated for 15 minutes with 1 mM dibutyryl cyclic AMP (DBcAMP) dissolved in water (Biolog, Bremen, Germany) and the second aliquot with water only.. 30 μL of 8- (2-Aminoethylamino)adenosine- 3', 5'- cyclic monophosphate (8-AEA-cAMP) crosslinked to agarose beads (Cedarlane Laboratories ltd, Canada, ON) were added to each of the two aliquots and incubated at 4°C for two hours. The agarose beads were then washed three times successively with equal volumes of phosphate-buffered saline (PBS) (pH 7,4) + 1% (v/v) Triton-X-100 + protease inhibitors and pelleted by centrifugation (2500xg) at 4°C. The agarose beads and 25 μL aliquots of each discarded wash and of the lysates were boiled for 5 minutes with Laemmli sample buffer and loaded onto 10% polyacrylamide gel and proteins were resolved by SDS-PAGE. Proteins were transferred from gels onto PVDF membranes (BioRAD, CA) by semi-dry electroporation (Hoefer SemiPhor, Pharmacia Biotech, San Francisco, CA, USA). The membranes were blocked in Tris-buffered saline with 0.2% (v/v) Tween-20 (TBST) + 5% (wt/volume) milk for 16 hrs. The blocked membranes were incubated for 2 hrs at room temperature with the primary antibodies: anti-Rluc antibodies (Mab4410 *versus* Rluc-F[1], Mab4400 *versus* Rluc-F[2], (Chemicon, Temecula, CA)), anti-dihydrofolate reductase antibodies (D1067 *versus* DHFR-F[1,2], D0942 *versus* DHFR-F[3] (Sigma-Aldrich, St-Louis, MO)) and anti-GFP antibody (A6455 *versus* both “Venus” YFP-F[1] and “Venus” YFP-F[2]); (Molecular probe, Eugene, OR, USA), 11814460001 *versus* “Venus” YFP-F[2]; (Roche, Basel, Switzerland)). The membranes were washed 3 times with TBST and incubated for 90 minutes at room temperature with the appropriate secondary antibody coupled to horseradish peroxidase: anti-rabbit horseradish peroxidase (HRP)-conjugated antibody (#7074; Cell Signaling technology, Danvers, MA, USA) for primary antibodies raised in rabbit (D1067, D0942, A6455) and anti-mouse horseradish peroxidase (HRP) conjugated antibody (#7076, Cell Signaling technology, Danvers, MA, USA) for primary antibodies raised in mouse (Mab4400, Mab4410 and 11814460001), and washed with TBST. The membranes were revealed with the western lightning chemiluminescence reagent plus (PerkinElmer, USA) substrate of HRP with Kodak Biomax XAR film over 30 seconds and 20 minutes exposure. Comparison of the three PCAs (fig. S10 B) showed that Bcy1 remained bound to cAMP resin while Tpk2 is found in the supernatant for both Rluc and DHFR PCAs whereas both the Bcy1 and Tpk2 subunits remained associated with cAMP resin in the case of the “Venus” YFP PCA. Thus, in contrast to the irreversible Venus YFP PCA and as already demonstrated for the Rluc PCA (*5*), the DHFR PCA is reversible.

2. Creation of universal DHFR PCA fragment templates and creation of homologous recombination cassettes. We set out to create two universal oligonucleotide cassettes encoding each complementary DHFR PCA fragment and two unique antibiotic resistance enzymes to allow for selection of haploid strains that have been successfully transformed and recombined with one or the other homologous recombination cassettes. The universal cassettes were constructed in three steps as follows: First, the ADH terminator (ADHterm) was PCR amplified from *S. cerevisiae* genomic DNA with a forward primer containing a BamHI restriction site prior to a 5' XbaI restriction site and a reverse primer containing a BglII 3' restriction site using the high fidelity Accuprime Pfx DNA polymerase (Invitrogen, Carlsbad, California). The BamHI/ BglII digested PCR product was subcloned into the multicloning site of the pAG25 and pAG32 plasmids (henceforth called pAG25ADHterm and pAG32 ADHterm) (*7*). Second, the DHFR PCA N-terminal (F[1,2]) and C-terminal (F[3]) fragments, each proceeded 5' by a sequence coding for a 10 amino acid $(Gly.Gly.Gly.Gly.Ser)_2$ flexible polypeptide linker (henceforth referred to as "linker") were PCR amplified from pMT3 mammalian expression vectors harboring these constructs (*3*). These PCRs were performed with a forward primer containing a 5' BamHI restriction site and reverse primer containing a 3' XbaI restriction site. The linker-F[1,2] was subcloned into pAG25ADHterm between BamHI and XbaI restriction sites 3' to the ADHterm sequence, creating the pAG25-linker-F[1,2]-ADHterm cassette. The linker-F[3] was subcloned into the multicloning site of pAG32ADHterm between BamHI and XbaI restriction sites creating the pAG32 linker-F[3]-ADHterm. Each of the plasmids used for subcloning already contained unique antibiotic resistance cassettes that in the resulting constructs are 3' to the ADHterm. Thus, the final DHFR PCA F[1,2] universal template consists of pAG25-linker-F[1,2]-ADHterm followed by TEF promoter, nourseothricin N-acetyl-transferase (NAT1) that confers resistance to nourseothricin and finally a TEF terminator. The final DHFR PCA F[3] universal template is the pAG32 linker-F[3]-ADHterm followed by TEF promoter, hygromycin B phosphotransferase that confers resistance to hygromycin B and TEF terminator. The resulting universal templates were used to create homologous recombination cassettes for each of 5,756 budding yeast genes by PCR using 5' and 3' oligonucleotides consisting of 40-nucleotide sequences homologous to the 3' end of each ORF (prior to the Stop codon) and a region approximately 20 nucleotides from the stop codon. Design of recombination cassette and diagnostic PCR primers are described below (Section 3).

3. Oligonucleotide design and synthesis. The oligonucleotides used for the 3'-tagging of each ORF with recombinant DHFR PCA cassettes and the oligonucleotides used to perform diagnostic confirmation of successful transformations were designed as follows and are available in tables S8 and S9. Coding and downstream sequences of yeast ORFs were downloaded from the Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/) on March, 2004. Homologous regions for the forward oligonucelotides were 40 nucleotides long and their sequences corresponded to a sequence 5' to the ORF stop codon. The following sequence was added to the end of this homologous sequence: GGCGGTGGCGGATCAGGAGGC, which anneals to the 3' end of the TEF terminator region of antibiotic resistance cassettes that are 3' to linker-F[1,2]-ADHterm in pAG25 and linker-F[3]-ADHterm in pAG32 as described in Section 2, creating a specific 61 bp PCR oligonucleotide for each of 5,756 ORFs. Homologous regions for the reverse oligonucleotides were 40 nucleotides long and contained the sequence of genomic

DNA immediately 3' to the stop codon of each ORF. The following sequence was added 5': TTCGACACTGGATGGCGGCGTTAG, which anneals to the sequence of the linker in both linker-F[1,2]-ADHterm in pAG25 and linker-F[3]-ADHterm in pAG32, creating a 64 bp oligonucleotide for each of 5,756 ORFs.

Diagnostic PCR oligonucleotides were designed to correspond to sequences of the non-coding strand of DNA in a region from 100 to 1,000 bp downstream of each ORF. Using custom made Perl scripts, these regions were searched for short sequences (18-25 nucleotides) with the following properties: have a melting temperatures from 58 to 62 degrees, have high GC content (> 50%), have a G or C at the 3' end for better annealing of the 3' end and optimized primer extension, have no complementary ends and do not contain palindromes. The forward diagnostic oligonucleotide consisted of a region common to both homologous recombination cassettes within the terminator of the antibiotic resistance marker.

The oligonucleotides were custom synthesized (IDT, Coralville, IA, USA). Primer mass was determined by mass spectrometry for quality control of the synthesis. Primers were delivered lyophilized on 15, 384-well format microtitre plates. The oligonucleotides were resuspended in sterile distilled water at a concentration of 133 μM in their indexed well positions and stored at -20°C. An aliquot of the forward and the reverse primers for generating each ORF-specific homologous recombination cassette were mixed in a 60 μl volume to final concentration of 5 μM of each primer. The primer mixtures were used for PCR amplification of the cassette and remaining volume was kept at -20°C.

4. Creation of homologous recombination cassettes

The linker-F[1,2]-ADHterm-TEFpromoter-NAT1-TEFterm in pAG25 and linker-F[3]-ADHterm-TEFpromoter-HPH-TEFterm in pAG32 were PCR amplified with the ORF-specific oligonucleotides (section 3) to create the specific homologous recombination cassettes in assigned positions of 384-well PCR plates. The 30 PCR cycles were carried out with the high fidelity Accuprime Pfx DNA polymerase, with an annealing temperature of 59°C for 30 seconds, a 3-minute elongation at 68°C, in a 25 μl total volume reaction. 3 μl aliquots of 40 randomly chosen PCR products on each 384-well plate were mixed with brilliant blue tainted glycerol, resolved by gel electrophoresis of the mixture on a 1% agarose gel stained with ethidium bromide and visualized under an UV light for quality control of the PCR amplification. The PCR products were directly used for creation of recombinant strains and the remaining reaction volume was stored at -20°C. Final homologous recombination cassettes are referred to henceforth as F[1,2]-NAT1 and F[3]-HPH.

5. Creation of recombinant strains

The strains BY4741 (*MAT***a** *his3Δ leu2Δ met15Δ ura3Δ*) and BY4742 (*MATα his3Δ leu2Δ lys2Δ ura3Δ*) were transformed as described in (*8*) with PCR products amplified from the templates described above to create specific homologous recombination DNA fragments (Section 4). The protocol was adapted to large-scale transformation in 96 well plates as follows: 8 μl of PCR product was mixed with 10 μl of chemically-competent yeast and mixed with 72 μl polyethyleneglycol (PEG) buffer and incubated for 30 minutes at room temperature. Cells were then heat shocked for 15 minutes at 42°C in a water bath. The transformation buffer was replaced with 250 μl YPD and cells were left to recover for 4 hours at 30°C and were then plated on antibiotic containing YPD agar plates and allowed to grow for 4 days at 30°C. In all cases the BY4741 (*MAT***a**) strain was transformed with the F[1,2]-NAT1 cassettes and BY4742 (*MAT*α) with the F[3]-HPH cassettes. The transformed strains were grown on YPD agar plates plus appropriate antibiotic agar plates (100 μg/mL nourseothricin for *MAT***a** transformed strains (WERNER BioAgents, Jena, Germany) or 250 μg/mL hygromycin B for *MAT*α transformed strains (Wisent Corporation, Quebec, Canada).

Identification of successfully recombinant clones was performed as follows: Putative recombinant clones were picked by hand, cell lysis was performed by heat treatment and confirmation of the correct location of a genome insertion was determined by PCR using the diagnostic primers described above (Section 3). The yeast lysates were placed at indexed positions of the 96-well PCR plates and mixed with the ORF specific and cassette specific diagnostic primers. Annealing of the diagnostic primers was performed at 56°C and 35 PCR cycles were carried out with the regular Taq DNA polymerase, with elongation cycles of 1.5 minutes at 72°C, in a 50 μl total reaction volume. Aliquots of 30 μl from each PCR product position within each 96-well plate were mixed with brilliant blue tainted glycerol and gel electrophoresed on ethidium bromide stained 1% agarose gels. Correct recombination was confirmed based on sizes of PCR products. The success rate for obtaining positive recombinants ranged between 30 and 90% of diagnosed colonies per 96-well plate. This process was repeated for up to six rounds of recombinant colony selection if recombinants were not found in the first round. Our efforts resulted in successful creation of 4,326 *MAT***a** strains harboring unique ORF-F[1,2]-NAT1 fusions and 4,804 *MAT*α harboring unique ORF-F[3]-HPH fusions. The confirmed strains were glycerol-stocked in pre-assigned positions of 96-well plates and stored at -80˚C with a total of 120 plates.

6. Optimization of DHFR PCA screening conditions We optimized the DHFR PCA by selecting a subset of 380 *MAT***a** strains with ORFs tagged with F[1,2] mated to 380 *MAT*α strains with ORFs tagged with F[3], for a total of 145,000 crosses (Section 5). These were selected based on the knowledge that the protein products of each ORF expressed as a fusion to F[1,2] should interact with the protein product of at least one ORF expressed as fusion to F[3] when complementary *MAT***a** and *MAT*α strains are mated and resulting diploids selected for growth in the presence of methotrexate (positive reconstitution of the methotrexate-insensitive DHFR PCA reporter). The known protein-protein interactions were obtained from the hand-curated MIPS database of protein complexes (http://mips.gsf.de/genre/proj/yeast/) as of 21 July 2005. We used this array to optimize the protocols for large scale, solid phase mating; selection of a methotrexate concentration that assures minimal background growth of negative control strains (Section 8) and maximizes growth of DHFR PCA-rescued diploid strains at different temperatures while minimizing the incubation time. Briefly, methotrexate concentrations were tested between 25 μg/mL and 225 μg/mL (in 25 μg increments) with an optimal concentration found at ~ 200 μg/mL. While higher concentrations provided more stringent selection, we found that methotrexate had limited solubility above this concentration. For all concentrations of methotrexate, we found that the optimal incubation time was 96 hours of growth at 30°C, which were then the conditions we used for the large-scale screen.

7. Test for detection of structural and topological organization of a protein complex

To test for the capacity of the DHFR PCA to provide structural and topological information, we performed a screen for interactions among the subunits of the well-characterized RNA polymerase II (RNA Pol II) complex structure, which has been determined at high resolution (PDB file 1I3Q) (*5*). We constructed homologous recombination cassettes for the 10 subunits (RPB2, RPB3, RPB5, RPB8, RPB9, RPB10, RPB11, RPC10, RPO21, RPO26) and created *MAT***a** and *MAT*α strains in which each of these genes are fused to the DHFR PCA fragments according to the methods described in Section 5. We used engineered heterodimerizing mutants of the SspB dimer from *Haemophilus influenzae* to serve as a positive control for the DHFR PCA and a negative control for interactions with RNA Pol II (*9*). Expression plasmids harboring these control proteins fused to the DHFR PCA fragments were generated as follows: The heterodimerizing SspB mutants, $\mathrm{SspB_{LSLA}}$ and $\mathrm{SspB_{YGMF}}$ were amplified by PCR and subcloned into p413-TEF-Zip-linker-DHFR F[3] and p413-TEF-Zip-linkers-DHFR F[1,2], replacing the leucine Zipper (Zip) sequence. We replaced the existing His3 metabolic enzyme selection marker cassette in p413TEF with nourseothricin N-acetyl-transferase (NAT1) or hygromycin B phosphotransferase (HPH) antibiotic selection genes. Specifically the cassette encoding (NAT1) from pAG25 was PCR amplified and subcloned between the sequence corresponding to the His3 complementing cassette of p413-TEF-$\mathrm{SspB_{LSLA}}$-linkers-DHFR F[1,2]. These plasmids were renamed p41NAT. The cassette encoding HPH gene from pAG32 was PCR amplified and subcloned between the sequence corresponding to the His3 gene cassette of p413-TEF-$\mathrm{SspB_{YGMF}}$-linker-DHFR F[3]. These plasmids were renamed p41HPH. Final products were transformed into *MAT***a** or *MAT*α (p41NAT-TEF-$\mathrm{SSpB_{LSLA}}$-linkers-DHFR F[1,2] and p41HPH-TEF-$\mathrm{SspB_{YGMF}}$-linker-DHFR F[3]) and respectively selected on YPD (+100 μg/mL nourseothricin) or YPD (+250 μg/mL hygromycin B).

The ten *MAT***a** strains harboring genomic fusions of DHFR F[1,2] fused to the ten RNA Pol II subunit genes or the control *MAT***a** expressing the $\mathrm{SspB_{LSLA}}$ mutant fused to F[1,2] were mated in a one-to-one matrix with the ten complementary *MAT*α strains harboring genomic fusions of DHFR F[3] fused to the ten RNA Pol II subunit genes or the *MAT*α control strain expressing $\mathrm{SspB_{YGMF}}$ mutant fused to F[3]. The diploid strains were grown on synthetic complete medium (SC) (-met, -lys, +methotrexate 200 μg/mL) to select for growth of methotrexate resistant diploid strains and incubated at 30°C for 120 hours (fig. S11). The plates were then photographed with a high-resolution (4 Mega pixel, Canon) digital camera and resulting images were converted to 8-bit grayscale images and analyzed with ImageJ 1.36b software (National Institutes of Health, USA), measuring pixel intensities of the colonies over a constant area. Results were corrected by subtraction of the pixel intensity of a region without colonies to remove any background intensity of the plate (fig. S10A, B). Based on a t-test for differences in growth, we detected 30 of the 50 possible interactions among the RNA Pol II subunits (60%), including homodimers while the control heterodimerizing mutants of SspB fused to mDHFR fragments showed no significant interactions with any of the ten RNA Pol II subunits (fig. S10C). The detected interactions are consistent with observations of physical proximity inside the RNA polymerase II complex (fig. S10D and Fig. 3).

8. Large scale PPI screen

To test pairwise protein-protein interactions between the two collections of *MAT***a** and *MAT*α strains, each of 3,247 *MAT***a** strains harboring ORF-F[1,2]-NAT1 fusions (henceforth the “baits”) were individually mated with each of 4,804 *MAT*α strains harboring ORF-F[3]-HPH fusions (henceforth “preys”) arrayed on agarose plates as described below (fig. S1). Diploids were subsequently selected and plated on medium containing methotrexate to select for positive DHFR PCA reconstitution. Although our screen is based on a selection assay and thus amenable to a split-pool screening strategy, we performed individual crosses to avoid the overrepresentation of highly abundant interacting proteins that would result from growth competition within pools of clones. The bait-by-query-array crosses were performed as follows: First, to generate the prey array, the *MAT*α query strains were printed from 96-well glycerol stock plates (Section 5) to 15 yeast peptone dextrose medium (YPD) agar plates with hygromycin B (+250 μg/mL hygromycin B) to a density of 384 colonies per plates using a 96 pin robotically manipulated pin tool (0.787 mm flat round-shaped pins, custom AFIX96FP3 BMP Multimek FP3N, V&P Scientific Inc., San Diego, CA) and allowed to grow for 24 hours at 30°C. The colonies were transferred to four YPD agar plates with hygromycin B (+250 μg/mL hygromycin B) using a 384 pin robotically manipulated pin tool (0.356 mm flat round-shaped pins, custom AFIX384FP8 BMP Multimek FP8N, V&P Scientific Inc., San Diego, CA) to a density of 1,536 spots/plate and allowed to grow for 24 hours at 30°C. Subsequently, colonies from the four *MAT*α plates were transferred onto individual YPD agar plates using a robotically manipulated 1,536 pin tool (0.229 mm flat round-shaped pins, custom AFIX1536FP9 BMP Multimek FP9N, V&P Scientific Inc., San Diego, CA), necessary for the mating procedure with individual *MAT***a** bait strains, resulting in *MAT*α query strain arrays at a density of 6,244 pin transfers per plate in which the 4,804 *MAT*α query strains form colonies at defined positions. In addition, positive and negative controls for correct transfer of cells and selection conditions were introduced at 48 positions on the array in cross-shaped patterns that traverse the entire array (See Fig. 1B) The positive control strains consisted of 24 *MAT***a**/*MAT*α diploids harboring DHFR PCA fusions of the MCK1 and CDC19 proteins that are known to interact and for which the interactions were confirmed by the DHFR PCA (MCK1-F[3] and CDC19-F[1,2]) and 24 negative control *MAT***a**/*MAT*α diploids for the pair of proteins CLN3 and CDC19, which have not been shown to interact by any method and show no growth in the DHFR PCA (CLN3-F[3] and CDC19-F[1,2]). Remaining positions on the 6,244-position array were empty; no cells having been transferred to these positions.

For each of the bait-by-prey array crosses (fig. S1), individual *MAT***a** bait strains were grown separately in liquid YPD with nourseothricin (+100 μg/mL nourseothricin) to saturation. The *MAT***a** bait strains saturated culture were concentrated by centrifugation and removal of approximately 60% of the supernatant. The *MAT***a** bait strain was printed on top of previously printed *MAT*α query strains on a YPD containing plates, using a robotically manipulated 1,536 pin tool (0.229 mm flat round-shaped pins, custom AFIX1536FP9 BMP Multimek FP9N, V&P Scientific Inc., San Diego, CA), and incubated at 30°C for 24 hours to allow for mating. The mated colonies were selected the next day by transferring all colonies onto YPD agar (+ 100 μg/mL nourseothricin, +250 μg/mL hygromycin B). The selected *MAT***a**/*MAT*α diploids were transferred the next day onto synthetic complete medium (SC) (-met, -lys, +methotrexate 200 μg/mL) to select for growth of methotrexate resistant diploid strains and incubated at 30°C for 90 hours. Individual plates were then photographed with a high-resolution (4 Mega pixel, Canon) digital camera and resulting images were analyzed and growing colonies quantitated using a shape recognition algorithm (Section 9). The diploid strains created were heterozygous for the two tagged genes, thus minimizing growth defects that may result from reduced activity or abundance of the modified gene products and reducing the potential for non-specific interactions among highly expressed proteins. Bait screens for which no significant growth above background was observed could be due to low expression of the proteins under the growth conditions, to interference of the mDHFR fragments with the folding or stability of bait and prey proteins, or simply to the absence of expression or interaction of the bait proteins with any of the query proteins under the conditions tested

9. Control experiments for spontaneous, interaction-independent DHFR protein fragment complementation. Large-scale PPI screens often contain a large fraction of false-positive interactions, either because of experimental noise or from artifacts specific to the detection method. The latter is the most problematic because false-positive interactions are often reproducible and have strong signal to background ratios. This problem therefore must be dealt with experimentally. Previous large-scale PPI studies (e.g., (*10-12*)) indeed reported that a number of proteins show promiscuous patterns of interactions. These interactions are typically arbitrarily eliminated from filtered datasets. The source of spurious results in PCA could be due to spontaneous complementation (folding) of the DHFR PCA fragments into active enzyme in the absence of any physical interaction between bait and prey proteins. Specifically, some proteins fused to a PCA fragment, (e.g. F[1,2]) could complement the other PCA fragment, F[3], without necessarily interacting with the protein to which this second fragment is fused, i.e. interacting with the F[3] fragment alone. In a screen such as ours, these proteins could promiscouously interact with many proteins and thus produce a large fraction of false positive interactions. To test for this possibility, we constructed a set of control baits that consist of the 10-amino acid linker (Gly.Gly.Gly.Gly.Ser)$_2$ fused to each of the complementary F[1,2] or F[3] DHFR fragments or F[1,2] or F[3] fragments alone. These control baits were used to challenge the prey array and a newly constructed *MAT***a** F[1,2] array created with the same procedure described in Section 7. The fragments have to be expressed from plasmids that harbor both linker-fragment fusions or fragments alone and contain resistance cassettes compatible with our screening strategy, transformed into complementary strains and mated against the individual arrays as described in section 8. The control constructs were generated by PCR amplifying linker-fragment fusions or fragments alone from the universal PCA template constructs (Section 2) constructed in pAG25 (linker-F[1,2]-ADHterm) and pAG32 (linker-F[3]-ADHterm). We used forward primers containing a 5' XbaI restriction site upstream of sequence coding for the linker sequence or individual DHFR PCA fragments and reverse primers coding for the 3' region of the individual fragments upstream of a 3' XhoI restriction site-coding sequence. These were used to subclone the XbaI- and XhoI-digested PCR products into the expression vector p413TEF between the XbaI and XhoI of the multicloning site 3' of the constitutive TEF promoter. The existing His2 metabolic enzyme section marker cassette in p413TEF needed to be replaced with NAT1 and HPH antibiotic selection genes to be consistent with our screening strategy as follows: The cassette encoding nourseothricin N-acetyl-transferase from pAG25 (*7*) flanked by the TEF promoter in 5' and terminator 3' was amplified with TEF promoter forward primer and TEF terminator reverse primer. These primers insert an NdeI restriction site into the 5' region of the PCR product and a NsiI restriction site in the 3' region. The resistance cassette was subcloned into the control expression plasmids between the NdeI restriction and NsiI restriction sites that flank the His3 complementing cassette. These plasmids are henceforth referred to as p41NAT for the NAT1 gene, conferring resistance to the antibiotic nourseothricin of transformed yeast. The control fragment plasmids expressing a F[3] fusion were further digested with BstXI and BsmI to subclone the HPH gene from pAG32 (*7*), located between BstXI and BsmI, encoding hygromycin B phosphotransferase, which confers resistance to the antibiotic hygromycin B of transformed yeast. The expression of the control fragments was confirmed by Western blot using a rabbit anti-DHFR polyclonal antibody that specifically recognizes an epitope in the N-terminal F[1,2] fragment (Sigma D1067, 1:6000 (Sigma-Aldrich, St-Louis, MO)) F[1,2] or an anti-DHFR polyclonal antibody that specifically recognizes an epitope in the C-terminal F[3] fragment (Sigma D0942, 1:5000 (Sigma-Aldrich, St-Louis, MO)) (Data not shown).

10. Data acquisition, colony quantification and documentation, and statistical analyses. Positive protein-protein interactions were interpreted from growth of the diploid colonies on methotrexate-containing medium as described above (Section 7). Complete acquisition and analysis of each plate proceeded as follows: First, images of the diploid methotrexate selection (*MAT***a** bait/*MAT*α prey) array plates were taken after 96 hours of growth (as described above, Section 7) with a 4.0 Mega pixel camera (Powershot A520, Canon). Plate images were saved in JPG format at a resolution of 180 dpi and a size of 2,272 × 1,704 pixels. Second, all of the 3,301 plates (3,247 plates for the 3,247 different baits, 48 repeated plates and 6 plates for the control experiments) were manually inspected during the image analysis step and positions too close to the plastic edges of the plate and therefore uninterpretable were eliminated by setting the intensities of the first and the last colonies on the first row of the array to 0. At this stage we eliminated 44 plates from the final analysis because they displayed growth of colonies at empty positions, likely resulting from grid misalignment or contamination. In order to accurately identify bait ORF/prey ORF coordinates on the array and to extract the intensity of the colonies, we developed image recognition routines from the Image Processing Toolbox of Matlab (The MathWorks, Natick, Massachusetts). Each plate contained a set of positive and negative control colonies as described in Section 7, arranged in parallel X-patterns across the plates (Fig. 1A), which allowed us to identify misalignment or mispositioning of the colony arrays on each plate due to variation in the robotic pinning process. The first step of the image analysis was to determine expected centers of the colonies arrayed in 96 columns and 64 rows. In order to adjust for variation in plating and possible rotation of an image during image acquisition we manually defined the coordinate center of a first and a last colony in a first row of the array and of the last colony of the first column using Matlab build-in tools for accepting user input. Using the coordinates of these centers the image was rotated so that each column and row of colonies lies on a straight line. Next, images were adjusted using the same parameters and centers of all colonies were calculated. For each subsequent plate, manual reselection of colony centers was performed for colonies or positions at the extremities of a plate if array coordinate position centers deviated significantly from those of an earlier plate. Image analysis was performed in the same order as the images were acquired and thus deviations in the positioning of plate images occurred among groups of plates and were easily identified. However, we usually found that series of images taken on the same day (~100 plates per day were processed) did not have to be readjusted. Results of positional array adjustment and detection of colony centers were manually checked for all images taken during the large-scale screening (3,307 plates in 6,144 position format plus spontaneous complementation control experiments (Section 8) and repeats of 48 bait/prey repeat plates.

While the procedure described above defines approximate colony centers, we used a modified version of a previously described algorithm to accurately locate them (*15*). The assumption of the algorithm is that pixels have higher intensity values at colony centers. Thus, the algorithm processes columns and rows of pixels one at a time and finds areas of maximal pixel intensity. Intersection of columns and rows where corresponding sets of pixels have the highest 75th percentile values are used to find approximate colony centers. We modified this algorithm in order to handle an array of 6,144 colonies per plate (96 well format was used in the original work). Next, we used an automated Matlab procedure to calculate colony areas and intensities on each plate as follows (fig. S1) (please refer to Matlab documentation for detailed description of functions used for the analysis): 1), images were corrected for possible non-uniform illumination as described in (http://www.mathworks.com/products/image/demos.html?file=/products/demos/shipping/images/ipexrice.html) and 2), small objects that correspond to gel background, bubbles, plate edges or other anomalies were removed using the imopen function with the disk morphological structuring excluding elements of a radius smaller than 2 pixels (radius of 4 pixels was used on plate edges). 3), images were converted into binary format using the im2bw function with a threshold calculated by the graythresh function. In this format, pixels that correspond to colonies were set to 1 and background pixels to 0. Thus, connected components of binary images with pixel values equal to 1 (calculated with bwlabel and regionprops functions) corresponded to colonies. Identity (row and column number on the plate array) of each colony was calculated by comparing approximate colony centers (calculated as described above) and centers of these connected regions. Intensity of each colony was then extracted from the original image. If there was no overlap between colonies (i.e. there were no colonies that touched each other) these 3 steps were sufficient for the analysis. However, in a small number of cases, large colonies overlapped between adjacent positions and had to be deconvoluted. One of the simplest ways of discriminating intersecting circular objects is to calculate the distance transform for a binary image followed by a watershed transform. The watershed transform finds "catchment basins" (which represent circular colonies) and "watershed ridge lines" (which correspond to an edge where colonies touch each other) in an image by treating it as a surface (see Image Analysis watershed segmentation demo for details http://www.mathworks.com/company/newsletters/news_notes/win02/watershed.html). As a result of the watershed transform, pixels that lie on the border between objects can be identified. These border pixels are set to zero on the original binary image, thus separating overlapping colonies and step 3 is repeated in order to analyze separated colonies. If the size of a connected region identified in step 3 was larger than expected for a single colony (on average, 400 pixels) or a connected region could not be matched to a position on the array (fusion of several colonies will result in centering of a connected region to deviate from an approximate colony center) the following steps were performed: 4), extraction of a rectangular subpart of an original image that fully contains a connected region that is suspected to contain fused colonies, 5), apply a distance transform. The distance transform calculates for each pixel its distance from the nearest nonzero pixel in a binary image. In the case of two intersecting circular objects it will produce a set of values with a maximum value at the centers of the objects. We found that calculating an accumulation array by the Circular Hough transform and superimposing its local maximum on a distance transform array further improved the detection of circular colonies. Hough transform is another method for detecting circular objects. It’s a voting procedure that assigns a value to a pixel if it can be a center of a circle. As a result, central pixels in circular objects receive higher values. So step 6), superimpose local maxima of the accumulation array of Circular Hough transform onto a matrix produced by step 5. 7), apply the watershed transform. 8), in cases where a number of objects that were separated using steps 4 to 5 was different from the number of possible centers detected by the Circular Hough transform, we performed a watershed transform on the original grey scale image that corresponds to the connected regions requiring deconvolution. Raw colony intensities are available on the Michnick Lab website (http://michnick.bcm.umontreal.ca/).

11. Analysis of colony intensity distributions and benchmarking.
Raw colony intensity (sum of pixel values from a grayscale) distribution was approximately log-normal and centered around 4,300 (fig. S3). This distribution showed a steep decrease in frequency at around 10,000, after which it showed an increase in frequency. This second distribution represents the population of diploid colonies able to grow on methotrexate. This is also where we saw a clear distinction between the positive (median: 38,361) and negative control intensity (median = 5,192) distributions. We therefore reasoned that the threshold above which we could infer a protein-protein interaction (PPI) should be located at around an intensity of 20,000 (see below). It is important to note that controls were diploid strains printed directly onto the plates and that these control strains did not go through the solid phase mating procedures. This explains why several colonies were smaller than the negative controls. These likely represented variability in diploid cell transfers.

In order to determine the accuracy of our data, we first had to eliminate data for ORFs that showed interactions with control fragments (spontaneous fragment complementation; Section 8). To do this, we calculated z-scores for control plates and used cutoffs for accepting an interaction based on a visual inspection of distributions of plate scores. Cutoffs were conservatively assigned based on visual inspection of the control plate distribution by identifying all colonies with larger intensities than the background distribution. We identified 344 such proteins (table S1), which are enriched for those associated with ribosomes ($P < 2\times10^{-65}$) (table S2) and are highly expressed compared to the proteome (median $\log_{10}(\text{Abund}_{sd})$ = 3.27 vs 2.28 for the proteome, Wilcoxon-rank test: $P < 2.2\times10^{-16}$; fig. S6). Similar highly abundant proteins are also often observed as false-positives in spurious interactions in affinity purifications in particular several ribosomal proteins and others like Cdc19p, Eno2p, Tef2p and Tef3p (*10*). Global analysis of the network created by these proteins revealed that they show highly structured and similar patterns of PPIs, suggesting that they could also be identified by computational analysis. Proteins that show correlated patterns of interaction with that of the controls but that were not identified in this control experiment could then be eliminated as well (see below). After eliminating positions from the plates corresponding to these proteins, we determined the threshold above which we could confidently infer PPIs as follows: First, we used the MIPS catalog of protein complexes (ftp://ftpmips.gsf.de/yeast/catalogues/complexcat/complexcat_data_18052006) as of 18 May, 2006 as a source of True-positive (TP) interactions (n = 11,005 among 1,236 proteins), 503 of which could be potentially detected in our screen (bait and prey strains exist in our collection for each protein and crossing of each strain resulted in growth of at least a colony, suggesting that the DHFR fragment-tagged ORFs are expressed and can interact with other proteins). This set of manually annotated MIPS complexes serves as a benchmarking standard for networks enriched for soluble proteins such as TAP-MS (*14*) and also in studies of helical membrane protein PINs (n = 79) (*15*). A true-negative (TN) set of PPI was obtained from (*14*) and consists of pairs of proteins that are part of distinct complexes and are expressed in different cellular compartments or have anticorrelated patterns of gene expression. This set contains 266,858 interactions, 6,377 of which could potentially be detected in our screen as described for the TPs. These two sets of PPIs allowed us to determine at what colony intensity we could confidently infer a PPI. In order to control for plate-to-plate variation in overall growth intensity and the non-random distribution of the number of interactions among plates, we combined two criteria (*16*) to determine a threshold above which to call the growth of a colony: The first criterion was the absolute intensity (a sum of pixel intensities on a grey-scale) of the colonies and the second, a score derived from the distribution of intensities on the plates:

$$\text{z-score}(x) = (x-\mu_x)/\sigma_x$$

where x is the intensity of a given colony, μ_x is the average intensity of the plate and σ_x is the standard-deviation of the mean. This allowed us to eliminate colonies that had high intensity values due to the high background growth of colonies on some plates. This combination allowed us to maximize coverage of true-positive PPI at a high Positive Predictive Value (PPV) comparable to small-scale experiments, where PPV is defined as the ratio of inferred True Positive interactions over the sum of the inferred True and False Positives. This was achieved with a pixel intensity of 23,000 and a z-score = 2.4. After removing false-positive interactions based on the control experiments (n = 344) and benchmarking on the MIPS gold standard, our final, high quality dataset includes data that reached a PPV score of 97.7%, implying that 97.7% of these interactions are predicted true-positives based on this high quality data. This cutoff results in data having precision comparable to all previous large-scale data sets including those that reach the same precision as small-scale protein interaction studies (Fig. 1C, fig. S5). At this cutoff and PPV score of 97.7% we observed 5,672 interactions**.** A further analysis revealed 83 highly connected proteins that mediate 2,902 interactions. Eight of these proteins constantly demonstrated a higher growth pattern than other *MAT*α strains (type 1). The remaining 75 proteins showed an unusually large number of common interactions. 23 proteins had a similar pattern of interactions to proteins that showed growth in the negative control experiments (type 2). Some of them, for example ARO8 and VAS1, were not among the 344 proteins identified as interacting with controls because their growth was just slightly below one of the two threshold cutoff values used to analyze control plates. The remaining 52 proteins showed a distinct pattern of common interactions (1,830 interactions) also typical of the 344 proteins that showed positive growing colonies in the negative control experiments. Given the ambiguity or control-like behavior of these results, all 2,902 interactions were removed from the final dataset. Our final, filtered dataset contained only 3 FP and 163 TP interactions. To determine the statistical significance of these results, we generated 10,000 random networks by randomly assigning the same number of interactions between the same set of proteins as in our final dataset to estimate the distribution of random scores, and a z-score for the observed score from this distribution. After applying a multi-step filtering procedure, the DHFR PCA network was derived from 24% of these bait screens. On average, a random network had 2.7 TP and 33.8 FP interactions. Our observed results are thus unlikely to have been found by chance alone (TP : $P = 0$, FP = $P < 10^{-7}$) and do not reflect a bias in the composition of our network with respect to the TP and TN data sets used here. A total of 3,113 colonies were found above the PPV threshold of 98.2%, which represents 2,770 unique interactions after subtracting interactions observed in both *MAT***a** bait (ORF1-F[1,2])/*MAT*α prey (ORF2-F[3]) and *MAT***a** (ORF2-F[1,2])/*MAT*α prey (ORF1-F[3]) (tables S3 & S4). 40% of the interactions that could have been detected in reversed crosses were detected, consistent with previous observations showing that interactions may be detected by the DHFR PCA when the individual proteins are fused to either F[1,2] or F[3] fragments, but not necessarily if the fragments are swapped between the two proteins (*17*). This proportion was 50% in the RNA pol II small-scale experiment (section 7), which suggests that we are near the upper bound with the large-scale screen. However, interactions that were observed in both directions (n = 343) can be considered as being of higher confidence given that their intensity scores on the plates is slightly higher on average (51, 216 versus 38,995), despite the fact the these two sets of proteins have only marginally different levels of protein abundances (median log10(Abund): 2.38 vs 2.32, P = 0.01). This is also reflected in the similarity of their Gene Ontologies. Pairs of proteins that were shown twice to interact during the screen have higher semantic similarity on average in terms of Biological Process, Molecular Function and Cellular Compartment than the same number of interactions taken randomly from the interactions detected in one direction only (BP : 3.45, BP random : 2.65, $P < 10^{-12}$; MF : 3.88, MF random : 3.12, $P < 10^{-5}$; CC: 2.76, CC random: 2.21 ; $P < 10^{-7}$).

PCA is efficiently able to detect interactions among membrane proteins with high specificity (*18, 19*), which contrasts with methods that have comprehensively studied protein interactions (*10, 11, 20-22*). About one quarter of all genes in most genomes contain putative transmembrane (TM) helices (*15*). They therefore likely represent an important fraction of the interactome and yet we have little knowledge of their patterns of interactions. The DHFR PCA network therefore represents an important steps in that direction. As a consequence of the difficulty to identify protein-protein interactions among membrane proteins, these are not well represented in public databases. Two papers established genome-wide protein interaction networks for membrane proteins: 1) Miller *et al.* (*23*) using the split-ubiquitin assay, and 2) Xia *et al.* (*15*), using computational prediction. These studies both acknowledged the challenge that the identification of protein-protein interactions among membrane proteins represents. Miller et al. identified 1,985 putative interactions among 536 proteins. Xia *et al.* predicted 4,145 interactions among 1,048 putative helical membrane proteins that they identified using computational methods.

As membrane proteins are less represented in high-quality datasets such as the MIPS (6% of proteins membrane proteins and 1% of interactions are among membrane proteins) which may limit our ability to identify false-positive interactions, we took special care in trying to identify potential spurious interactions involving membrane proteins. First, the control experiment with the fragments and the linkers alone revealed 19 (out of 344) putative helical membrane proteins that show spontaneous fragment complementation. Further, we eliminate 83 proteins that showed similar patterns of interactions, which is a characteristic of proteins showing spontaneous complementation with other proteins. This analysis further identified 45 membrane proteins that are likely to mediate spontaneous complementation of the DFHR fragments and these were removed from the dataset. In the final network, 232 of the putative membrane proteins identified by (*15*) show interactions in our PCA screen out of 1,124 proteins, which represents a slight enrichment compared to the entire genome. We identified 2,770 high quality interactions among 1,124 proteins in our screen. The average degree of a protein (number of interactions) is therefore ~ 2.5. Our network contains 232 helical membrane proteins that make 662 protein interactions among them, for an average degree of ~2.8. This is not much higher than for the network as a whole. This may represent an increased power of the PCA to detect pairwise interactions among membrane proteins, which are less spatially constrained than proteins that are member of large complexes. In order to assess the quality of the fraction of the DHFR PCA network that represents the among helical membrane protein interactions, we examined the overlap between our final dataset and that of the two previous membrane protein PINs. First, we found that we detected 51 of the 662 interactions predicted by (*15*). Randomizations (10,000) of our network reveals that we expect an overlap of only 5 interactions between these two datasets by chance alone, which represents an enrichment of 10 fold over the random expectation ($P < 10^{-94}$). Similarly, 27 of the our 662 interactions were also identified using the split-ubiquitin method (*23*). Only 1.9 interactions are expected to be in common between these two datasets, which represents an enrichment of 15 fold over the random expectation ($P < 10^{-75}$). Finally, we found that while controlling for the sharing of cellular compartment of interacting protein pairs, interacting protein pairs remain significantly enriched for semantic scores of Molecular Function and Biological processes (i.e. similarity of gene ontology categories), and this for the whole set of protein interactions (MF: 3.76, MF random: 2.48, $P < 10^{-74}$; BP: 3.34, BP random: 2.14, $P < 10^{-144}$) and for the set of helical membrane proteins interactions (MF: 3.89, MF random: 2.84, $P < 10^{-18}$; BP: 2.86 , BP random: 2.15, $P < 10^{-15}$).

12. Analysis of high-quality PPI in comparison to protein abundance, gene ontology enrichment and three-dimensional structures.

12i. Analysis of PPI versus protein abundance Protein abundance for yeast grown under the same conditions (SC medium) as used in the DHFR PCA screen or YEPD and based on FACS analysis of GFP- tagged proteins was obtained from (*24*).

12ii. Analysis of Gene Ontology enrichment. We examined whether the final set of proteins constituting the DHFR PCA network contained an overrepresentation of proteins involved in specific biological processes, with certain molecular functions or that localize to specific cell compartments**.** Gene Ontology enrichments were calculated using the method implemented in GOstat (R library) (*25, 26*). This approach utilizes a conditional hypergeometric algorithm that considers the hierarchical relationships of gene ontology definitions to decorrelate the results. More precisely, due to their hierarchical organization—a GO term inherits all annotations from its more specific descendants. Gene Ontology categories are not exclusive and are thus locally correlated. This method considers this organization to limit the redundancies in the results. The algorithm considers more specific to more general terms. When testing the significant enrichment of a GO term, it removes genes that are annotated to a significantly enriched node from all its ancestors (more general term). The universe of genes (reference list to which the network proteins are compared) used for the comparisons was the entire genome, as we aimed at identifying what categories of proteins were overrepresented relative to the entire proteome. A conservative *P-value* cutoff of 0.00001 was used for these analyses.

12iii. **Correlation of gene expression profiles.** Correlation of gene expression between pairs of ORFs (Pearson correlation) was calculated from gene expression profiles from more than 1,000 expression profiles compiled in (*27*). The distribution of correlation coefficients for random networks of the same size and degree distribution were estimated from 1,000 random networks.

12iv. Structural analysis of PPI Structural domain annotations for all *S. cerevisiae* proteins were obtained from SGD in September 2007. The identity of domains known to mediate PPI were obtained from (*28*). These data are derived from the mapping of protein domains onto 3D structures of resolved protein complexes deposited in the Protein Data Bank by (*29, 30*). Domains mediating PPIs are those that are found at the interface of interacting proteins. The fraction of protein pairs in the DHFR PCA network that each have a domain that is known to mediate PPIs was estimated by combining these data with the domain identity obtained from SGD. The random expectation of the fraction of interacting proteins pairs that have one domain each that mediate PPI was estimated by generating random networks, as described above. Distances among C-termini of known protein complexes were determined as follows: The identity of yeast proteins that have homologs in the Protein Data Bank (PDB) was obtained from SGD. On June 2007, 46,320 PDB files corresponding to biological units of cellular organisms were obtained from PDB, among which 13,966 contained yeast homologs. The first biological unit was used (.pdb1) if more were available. We extracted the distances between the C-termini of all pairs of chains within each complex. We then went through each complex and determined whether: 1), they contain more than one chain homologous to any of the proteins that are part of the DHFR PCA network, 2), we recorded all of the possible pairs of interactions within those complexes and 3), we merged all the interactions together and if a pair of chains was recorded more than once, we kept the one with the shortest distance between the C-temini of these two proteins. Finally, we examined whether this interaction was seen or not in the DHFR PCA screen. We considered interactions as not being seen only if it could have been detected, i.e. the pair of proteins showed at least one interaction as bait and prey or as prey and baits. This left us with 175 interactions from 129 distinct PDB entries.

12v. Membrane topology and PPI determined by DHFR PCA. Protein membrane topology was obtained from (*31*). We reasoned that the C-termini of interacting proteins have to be oriented into the same cellular compartment in order for DHFR PCA to occur; for instance both in the cytosol or both in the lumen of the endoplasmic reticulum. Kim *et al.* (*31*) established the location of the C-termini of 546 proteins of which 448 have their C-termini oriented towards the cytosol. We examined all possible pairs of proteins among these 546 proteins and determined which of these are localized in at least one common cellular compartment using microscopic evidence for yeast proteins fused to green fluorescent protein (*32*). It is important to note that we don't know if any of these pairs should actually show an interaction. For these possible interactions, we counted how many of each type we detected and did not detect, considering only those for which interactions could have been observed as described above (i.e. strains exist in which both proteins are fused to one or the other DHFR PCA fragments and show interactions).

13. Comparison of DHFR PCA to previously determined PPI. In order to examine whether the DHFR PCA interactions had been reported previously, we obtained the following databases: Biogrid (www.thebiogrid.org/, version 2.0.29), mips-MPact (http://mips.gsf.de/genre/proj/mpact, version 18052006) and DIP (http://dip.doe-mbi.ucla.edu/, June 2007). Entries in Biogrid with experimental systems defined as: "Synthetic Lethality" , "Dosage Rescue", "Synthetic Growth Defect", "Synthetic Rescue", "Epistatic MiniArray Profile", "Dosage Lethality", "Phenotypic Enhancement", "Phenotypic Suppression" and "Dosage Growth Defect" were not considered. Mpact entries were considered only if they had the tag "902.01.01.02.01", which indicates physical interactions. We separately considered the combined TAP-MS data from (*14*) (data used were those with Purification Enrichment scores of 3.19 and above as defined in (*14*)), as it overlaps considerably with what has been deposited in Biogrid by (*11*) and (*10*). We considered only one citation for an interaction reported in Collins *et al.* (2007), even if it had been reported by one of, or both of the two original studies. Finally, we considered only interactions associated with PubMed ids, because these can be tracked to their original experimental evidence. We considered different ids as being independent evidence. Although this may inflate the confidence in some interactions, it should not affect the identification of new interactions as described in our screen.

14. Overlap with previous large-scale studies. In order to calculate the overlap between our screen and previous results derived from large-scale experiments and a catalogue of manually annotated complexes we computed a number of interactions common to our screen and each of the datasets described below. We first calculated the overlap between DHFR PCA PPI and all interactions reported in a previous study (figs. S7 & S8; numbers in circle indicate the total number of interactions detected in a particular dataset (since affinity-purification based methods cannot detect homomeric interactions, only heteromeric interactions were used for the analysis)). None of the datasets reports a complete set of interactions between all yeast proteins and thus low overlap between different data sets may be because different datasets cover different sets of yeast proteins. Therefore, we performed a normalization of the number of interactions based on how many proteins are in common between two datasets and how many same interactions could be detected (fig. S8). Numbers in the PCA circles indicate how many interactions reported in a particular dataset could be detected by our screen (an interaction can be detected only if one of interacting partners showed a signal as bait and another as prey or vice-versa in our final network). Numbers in circles that correspond to a reference dataset indicate how many interactions from our final network could be detected by a reference dataset. When only interactions that could be detected by both experiments are considered, the overlap between DHFR PCA and reference datasets reaches as high as 50%. Reference datasets and criteria for normalization are described below:

1), MIPS catalogue of manually annotated complexes. Downloaded from ftp://ftpmips.gsf.de/yeast/catalogues/complexcat/complexcat_data_18052006. Complexes detected by large-scale experiments were filtered out from this file and interactions were assigned between all proteins that belong to the same complex. An interaction is considered to be possible if both interacting partners are present in the MIPS catalogue.
2), Krogan *et al.* (*10*) The core dataset was obtained from supplementary table 7, that lists successfully identified baits and preys obtained from supplementary tables 2 and 3. An interaction is considered to be possible only if one of the proteins is present in the baits list and another is in the preys list. We don’t consider a co-occurrence of both proteins in the prey list since the core dataset of this study contained only bait-prey pairs.
3), Gavin *et al.* (*11*) The network of interactions was obtained as deposited in Biogrid. This study used a statistical framework for deriving a high confidence set of interactions that makes possible interactions between two prey proteins. Therefore for normalization, we considered an interaction to be possible if for a pair of interaction partners, a bait-prey or prey-prey pair exists in the raw purification data (downloaded from http://yeast-complexes.embl.de).
4) Collins *et al.* (*14*) We used high confidence data with a PE score cutoff of 3.19. Normalization was performed as described above for Gavin *et al.* using a combination of both Krogan and Gavin raw datasets.

5), Ito *et al.*, Uetz *et al.* (*22*) Interactions detected by yeast two-hybrid assays. Interaction was considered possible if one of the interacting partners is among proteins that showed an interaction when tagged with a binding domain and another is among proteins that showed interactions when tagged with an activation domain.
6), Miller et al. (*23*) Interactions tested using the split-ubiquitin reporter. The data were extracted from supplementary Table 1. An interaction is considered possible if a corresponding pair of Cub-PLV and NubG ORF is present in the dataset.

15. Clustering of high confidence interactions. 2,534 heteromeric high confidence interactions were clustered as described in (*33*). For clustering purposes and during calculation of the association matrix, a value for self-association of a protein was set to 1. Hence, observed homomeric interactions could not be distinguished from those used for clustering purposes. Since two proteins that belong to the same complex may not be interacting but rather kept together by a common interacting partner, an association matrix based on the number of links that connect two proteins can be used for clustering (*34*). We found that for the sparse PPI matrix derived from DHFR PCA data, this procedure resulted in a tight and natural clustering of interactions among subunits of known complexes. Briefly, the network was organized into an association matrix with entries for pairs of proteins that range between 0 and 1. Values were calculated as $1/d^2$, where d is the shortest path in the network between these two proteins. A hierarchical agglomerative average-

linkage clustering with the uncentered correlation coefficient as the distance matrix was then applied to the association matrix (*34*). The data were then visualized using iVici (File S1, http://michnick.bcm.umontreal.ca/ivici/) (*35*). Only direct interactions are shown on a complete map (Fig. 4). On the insets, direct interactions are bright red, while indirect interactions (2 or 3 links between two protein) are shown as two consecutively darker shades of red, respectively. The clustered network is available in Supplementary File S1. This file can be opened and visualized using iVici (*35*), a platform independent software available at **(**http://michnick.bcm.umontreal.ca/ivici/).

16. GO enrichment. The GO slim map was obtained from SGD 17 February 2007. For every pair of annotation terms associated with a biological process, cellular compartment and molecular function we calculated enrichment/depletion in the number of interactions in our high confidence dataset compared to the number expected by chance. First, we calculated the number of interactions that are detected between proteins associated with a specific pair of GO terms. Next, we constructed 10,000 randomized networks of interactions between the same set of proteins as in our high confidence network and with the same number of interactions per proteins. The randomized networks contain only interactions that could be detected in our screen (as described for the normalized overlap calculations). For each randomized network, a number of interactions between proteins associated with a specific pair of terms was calculated and a z-score was derived by comparison of these numbers with a corresponding number of interactions detected by our analysis. High z-scores correspond to significant enrichment in interactions comparing to random while negative z-scores values correspond to significant depletion.

For pairs of interacting proteins that lack a common GO slim annotation term we calculated a GO semantic score, which is another measure of a functional relationship between proteins. The semantic score takes into account a specificity and hierarchy of a parent term that is common to a pair of proteins. Thus, pairs of proteins that are more closely related have higher scores (*36*). The results were compared with semantic scores calculated for random networks generated as described above.

List of Supplementary Tables

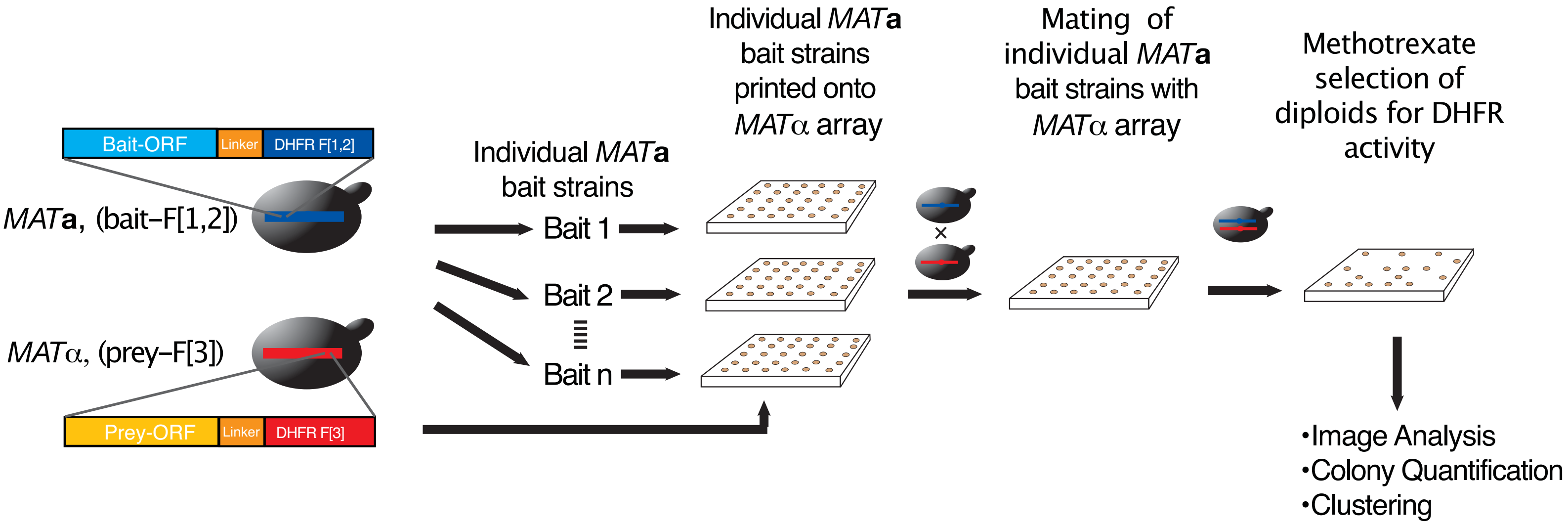
Table S1: List of ORFs successfully tagged.
Table S2: List of control proteins
Table S3: The DHFR PCA network.
Table S4: Interactions and their associated intensities, z-score and PPV values, for PPV values above 80%. *MAT***a** and *MAT*α types are described in section 11.
Table S5: Gene Ontology Enrichments of DHFR PCA proteins.
Table S6: Gene Ontology Enrichments of combined TAP-MS from Collins *et al.* 2007.
Table S7: Inter-MIPS complexes DHFR PCA interactions and their semantic scores
Table S8: Oligonucleotides for the amplification of the tagging cassettes for homologous recombination.
Table S9: Oligonucleotides for confirmation of homologous recombination by diagnostic PCR.
File S1: Clustered DHFR PCA network as described in Fig. 4, to be opened and visualized using iVici, downloadable at (http://michnick.bcm.umontreal.ca/ivici/).

References to supplementary material

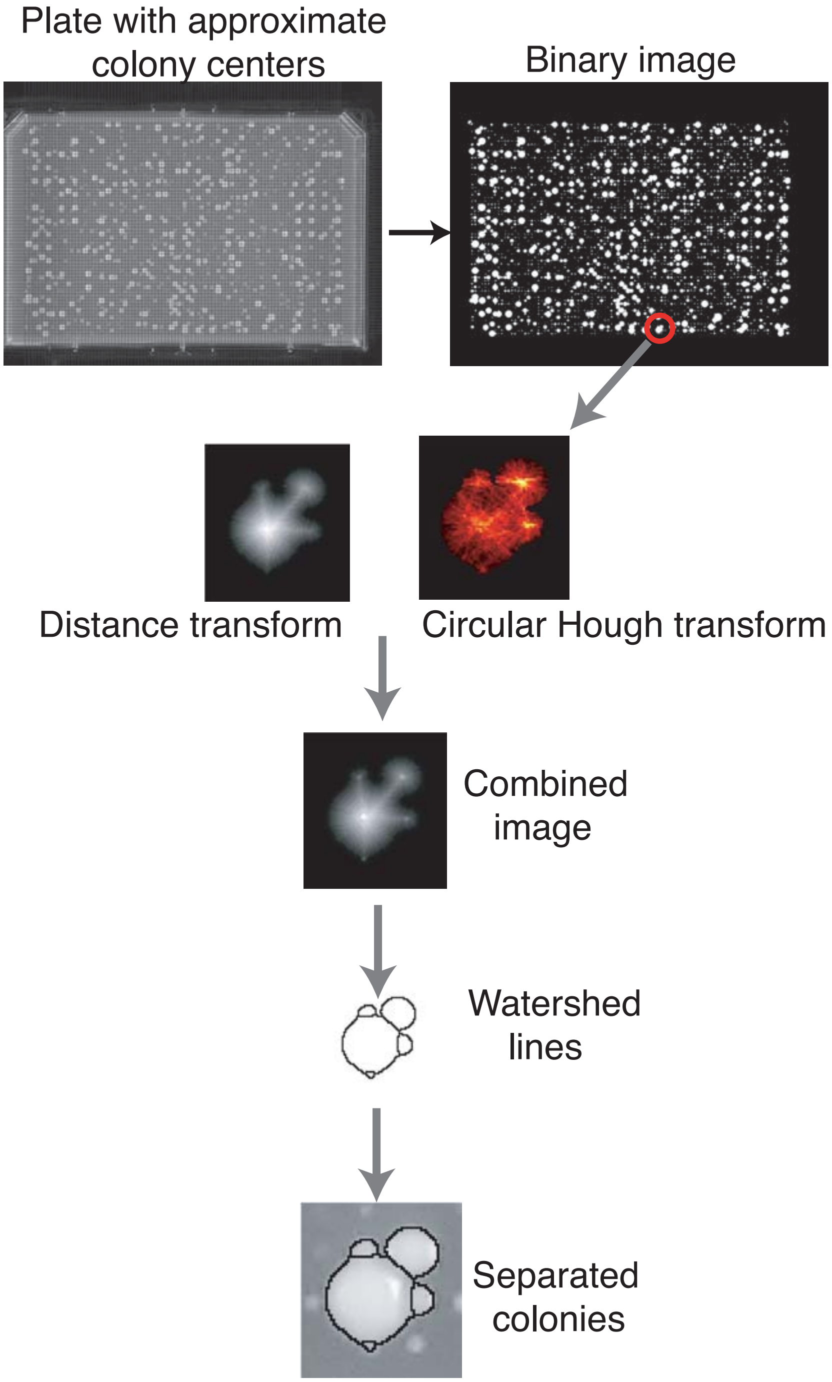
1. J. N. Pelletier, F. X. Campbell-Valois, S. W. Michnick, *Proc Natl Acad Sci U S A* **95**, 12141 (1998).
2. R. Subramaniam, D. Desveaux, C. Spickler, S. W. Michnick, N. Brisson, *Nat Biotechnol* **19**, 769 (2001).
3. I. Remy, S. W. Michnick, *Proc Natl Acad Sci U S A* **96**, 5394 (1999).
4. I. Remy, S. W. Michnick, *Nat Methods* **3**, 977 (2006).
5. E. Stefan *et al.*, *Proc Natl Acad Sci U S A* **104**, 16916 (2007).
6. J. Ramseyer, C. B. Kanstein, G. M. Walton, G. Gill, *Biochim Biophys Acta* **446**, 358 (1976).
7. A. L. Goldstein, J. H. McCusker, *Yeast (Chichester, England)* **15**, 1541 (1999).
8. R. D. Gietz, R. A. Woods, *Methods Enzymol* **350**, 87 (2002).
9. D. N. Bolon, R. A. Grant, T. A. Baker, R. T. Sauer, *Proc Natl Acad Sci U S A* **102**, 12724 (2005).
10. N. J. Krogan *et al.*, *Nature* **440**, 637 (2006).
11. A. C. Gavin *et al.*, *Nature* **440**, 631 (2006).
12. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
13. N. A. Shah, R. J. Laws, B. Wardman, L. P. Zhao, J. L. t. Hartman, *BMC Syst Biol* **1**, 3 (2007).
14. S. R. Collins *et al.*, *Mol Cell Proteomics* **6**, 439 (2007).
15. Y. Xia, L. J. Lu, M. Gerstein, *J Mol Biol* **357**, 339 (2006).
16. C. Brideau, B. Gunter, B. Pikounis, A. Liaw, *J Biomol Screen* **8**, 634 (2003).
17. I. Remy, S. W. Michnick, *Proc Natl Acad Sci U S A* **98**, 7678 (2001).
18. R. Benton, S. Sachse, S. W. Michnick, L. B. Vosshall, *PLoS Biol* **4**, e20 (2006).
19. I. Remy, I. A. Wilson, S. W. Michnick, *Science* **283**, 990 (1999).
20. T. Ito *et al.*, *Proc Natl Acad Sci U S A* **98**, 4569 (2001).
21. T. Ito *et al.*, *Proc Natl Acad Sci U S A* **97**, 1143 (2000).
22. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
23. J. P. Miller *et al.*, *Proc Natl Acad Sci U S A* **102**, 12123 (2005).
24. J. R. Newman *et al.*, *Nature* **441**, 840 (2006).
25. S. Falcon, R. Gentleman, *Bioinformatics* **23**, 257 (2007).
26. R. Ihaka, R. Gentleman, *Journal of Computational and Graphical Statistics* **5**, 299 (1996).
27. J. Ihmels, S. Bergmann, N. Barkai, *Bioinformatics* **20**, 1993 (2004).
28. Z. Itzhaki, E. Akiva, Y. Altuvia, H. Margalit, *Genome biology* **7**, R125 (2006).
29. A. Stein, R. B. Russell, P. Aloy, *Nucleic Acids Res* **33**, D413 (2005).
30. R. D. Finn, M. Marshall, A. Bateman, *Bioinformatics* **21**, 410 (2005).
31. H. Kim, K. Melen, M. Osterberg, G. von Heijne, *Proc Natl Acad Sci U S A* **103**, 11142 (2006).
32. W. K. Huh *et al.*, *Nature* **425**, 686 (2003).
33. A. W. Rives, T. Galitski, *Proc Natl Acad Sci U S A* **100**, 1128 (2003).
34. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (1998).

35. K. Tarassov, S. W. Michnick, *Genome biology* **6**, R115 (2005).

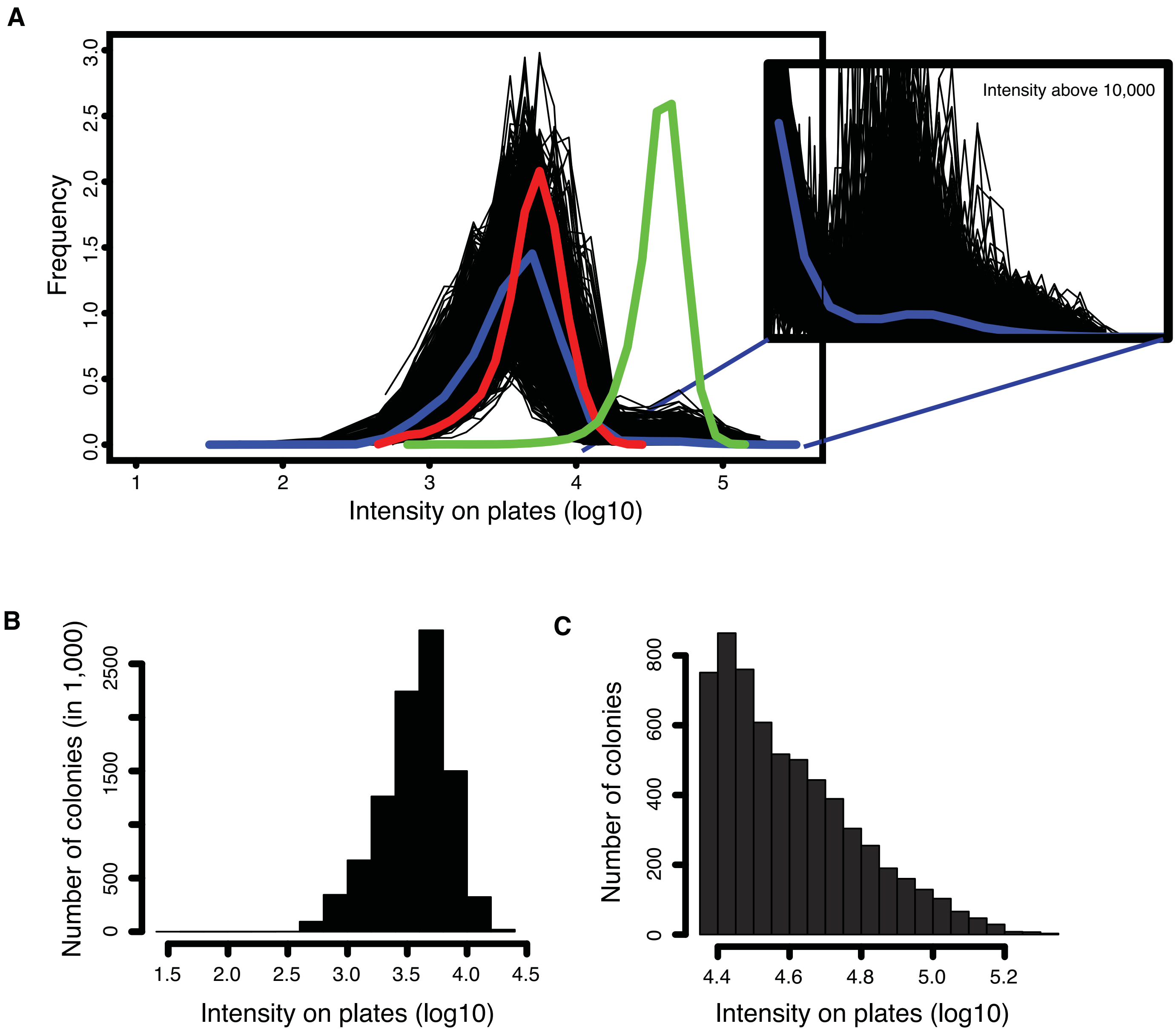
36. P. W. Lord, R. D. Stevens, A. Brass, C. A. Goble, *Bioinformatics* **19**, 1275 (2003).


Bait-ORF
Linker
DHFR F[1,2]
*MAT***a**, (bait–F[1,2])
*MAT*α, (prey–F[3])
Prey-ORF
Linker
DHFR F[3]
Individual *MAT***a** bait strains
Bait 1
Bait 2
Bait n
Individual *MAT***a** bait strains printed onto *MAT*α array
×
Mating of individual *MAT***a** bait strains with *MAT*α array
Methotrexate selection of diploids for DHFR activity
•Image Analysis
•Colony Quantification
•Clustering

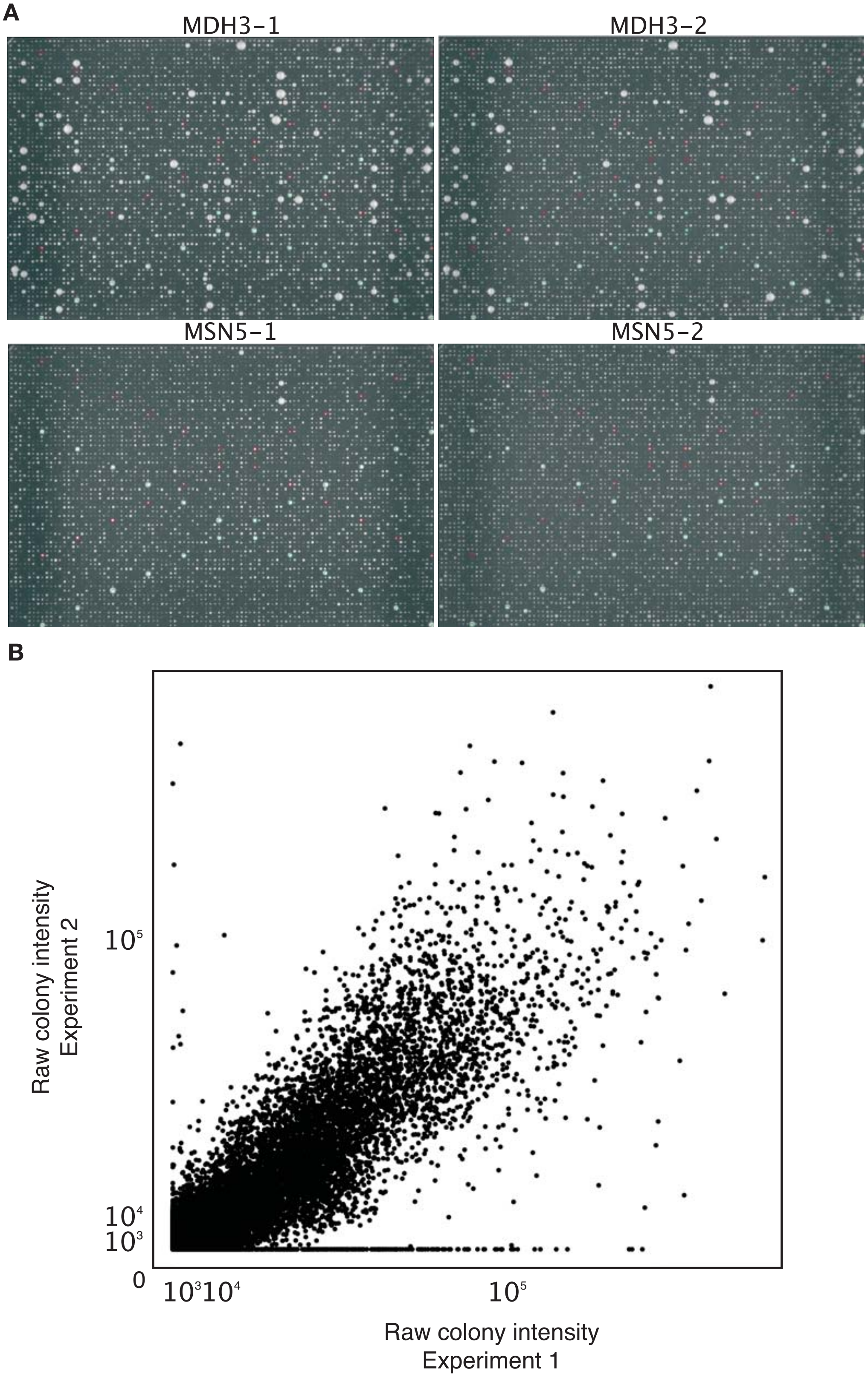
Supplementary figure 1: *In vivo* PCA screen of the yeast protein interaction network. Strategy for single bait versus prey array screening of the yeast PIN by DHFR PCA.



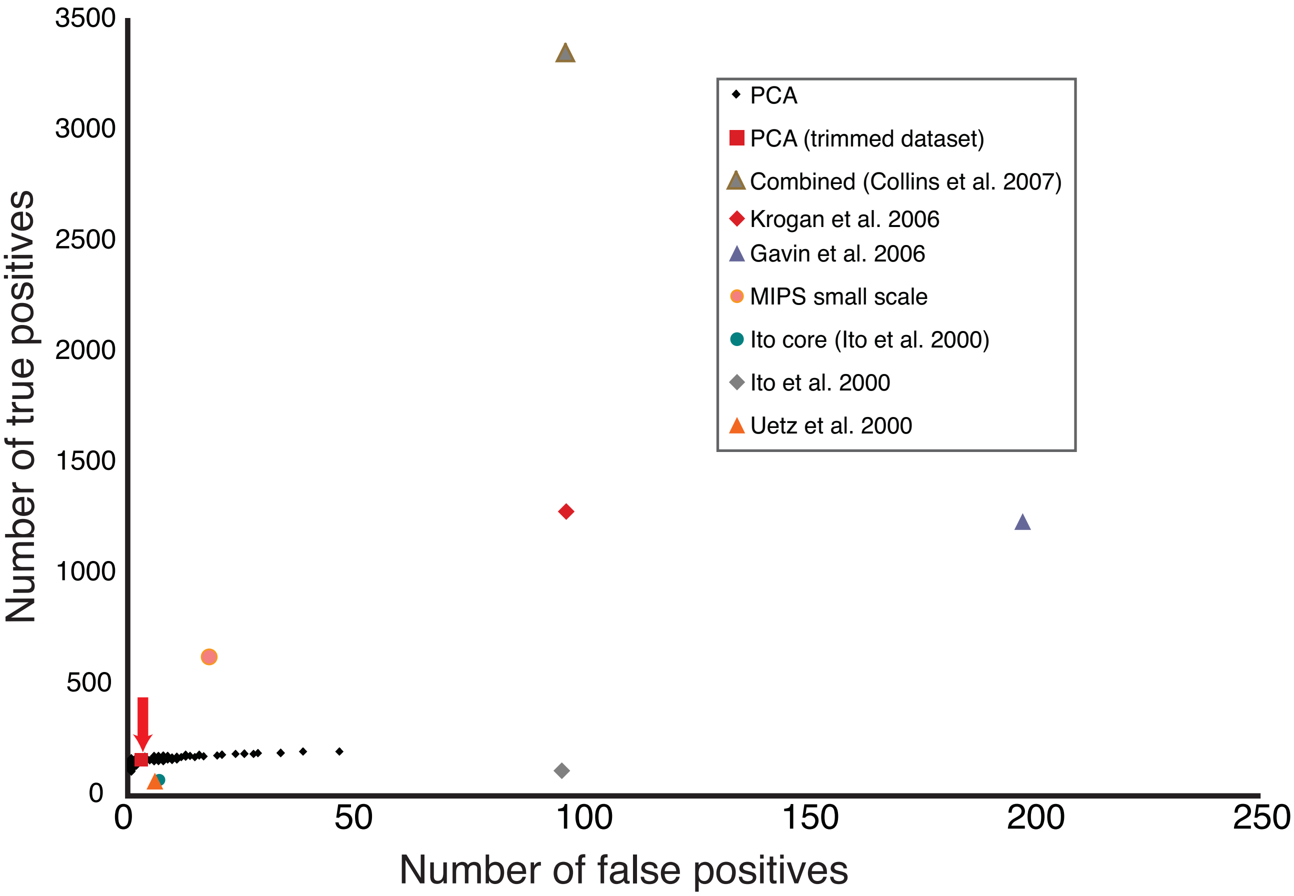
Supplementary figure 2: Automated extraction of colony intensities on plates. The DHFR PCA results were inferred from the growth of diploid colonies on plates containing methotrexate. Images of the plates were taken after a 90-hour growth period with a 4.0 Mega pixel camera (Powershot A520, Canon). Plate images were saved in JPG format at a resolution of 180 dpi and a size of 2,272 × 1,704 pixels. In order to extract the intensity of the colonies, we used available image recognition routines available in the Matlab image analysis toolbox and we modified parameters for it to be able to differentiate colonies that are in proximity to each other. The quality of the position of the grid and the recognition algorithm was examined through visual inspection of all plates.



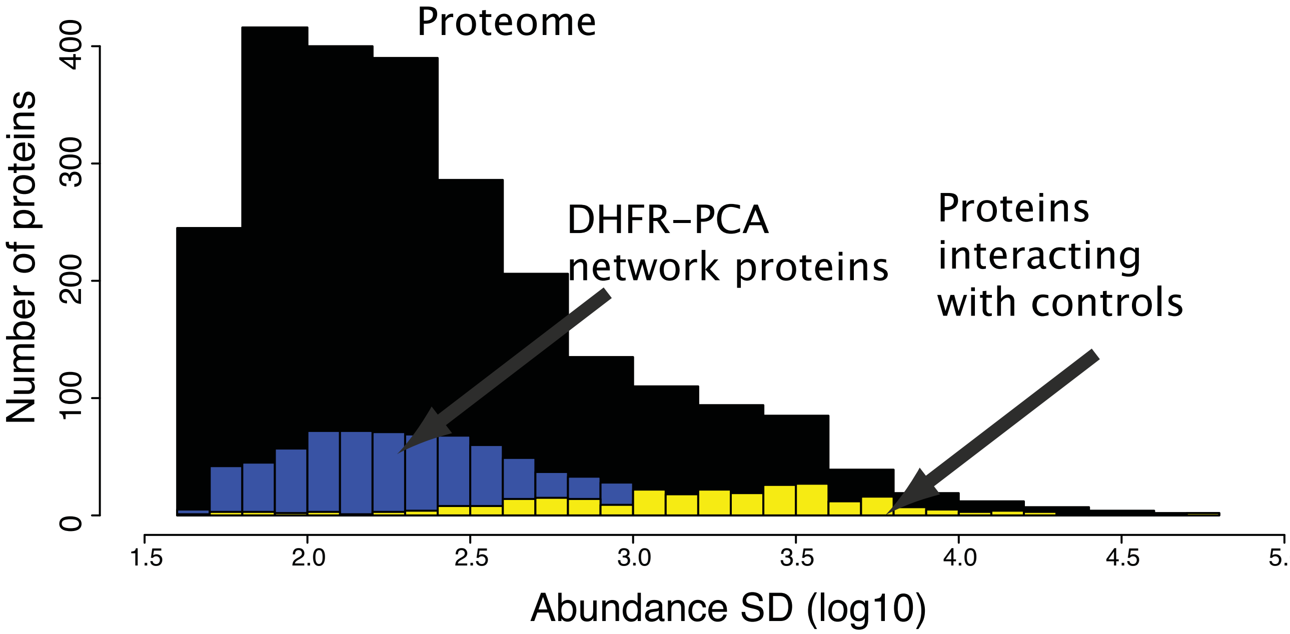
Supplementary figure 3: Distribution of colony intensities on plates. (A) Raw intensities prior to filtering. Black lines represent the intensity distributions on individual plates. The blue line represents the distribution of colony intensities across the entire experiment. The red and green lines represent the intensity distribution of the negative and positive controls respectively. The second panel shows the distribution of colony intensities above 10,000 to illustrate the growth of the methotrexate resistant diploid strains. (B) Intensities of colonies below the threshold. (C) Intensities of colonies above the threshold after filtering positions corresponding to baits and preys that interact with the control fragments.



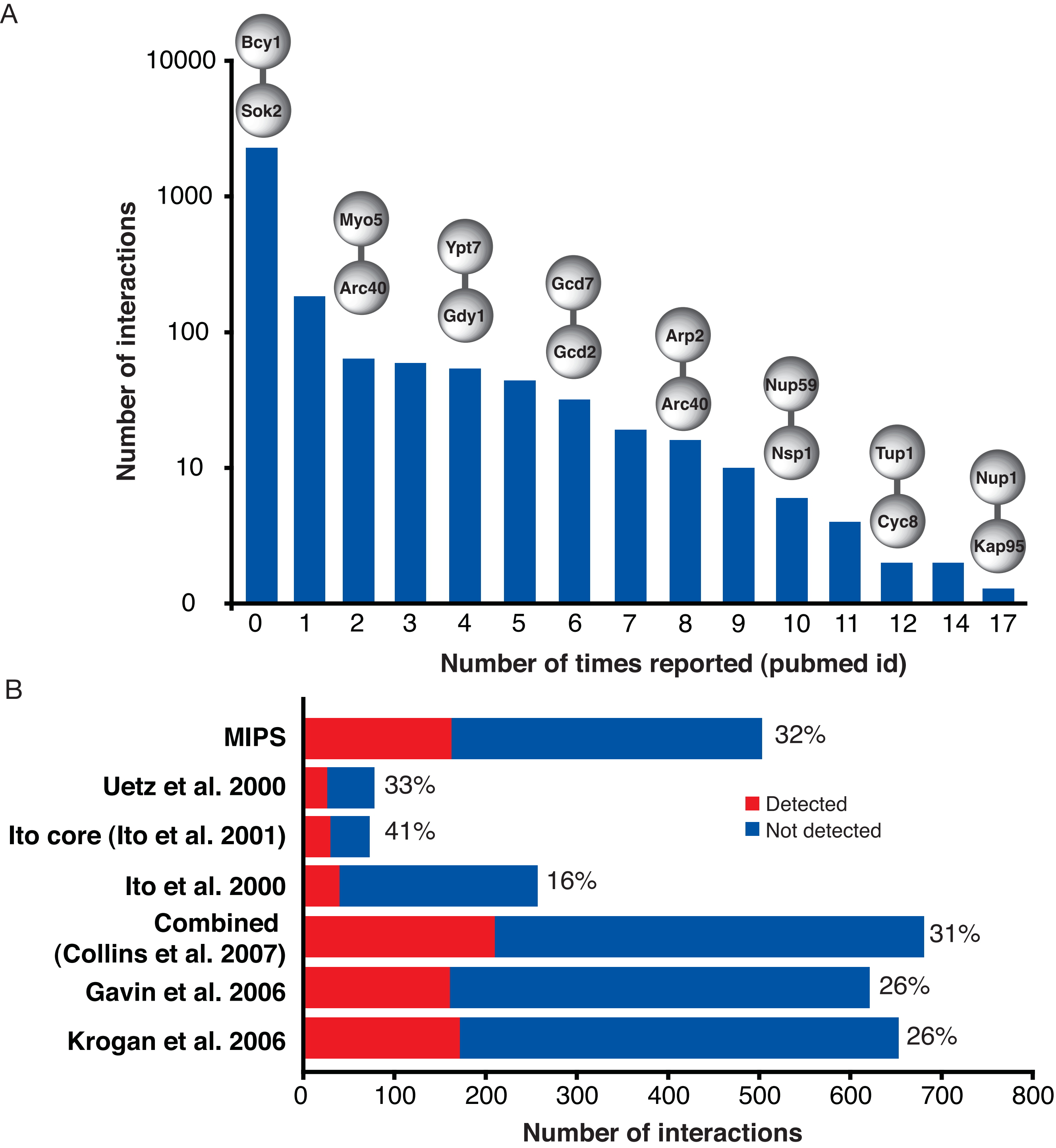
Supplementary figure 4: Reproducibility of the screening process. In order to evaluate the reproducibility of the screening process, we repeated the screen for 48 baits selected for representing the distribution of number of colonies growing above background in the first screen. (A) Example of plates that were repeated. Top row, MDH3 and bottom row, MSN5 used as baits. Green and red circles represent respectively positive and negative controls. (B) Raw colony intensity of experiment two plotted against the raw colony intensity of experiment one (Pearson correlation: 0.86, $P < 2.2e-16$).



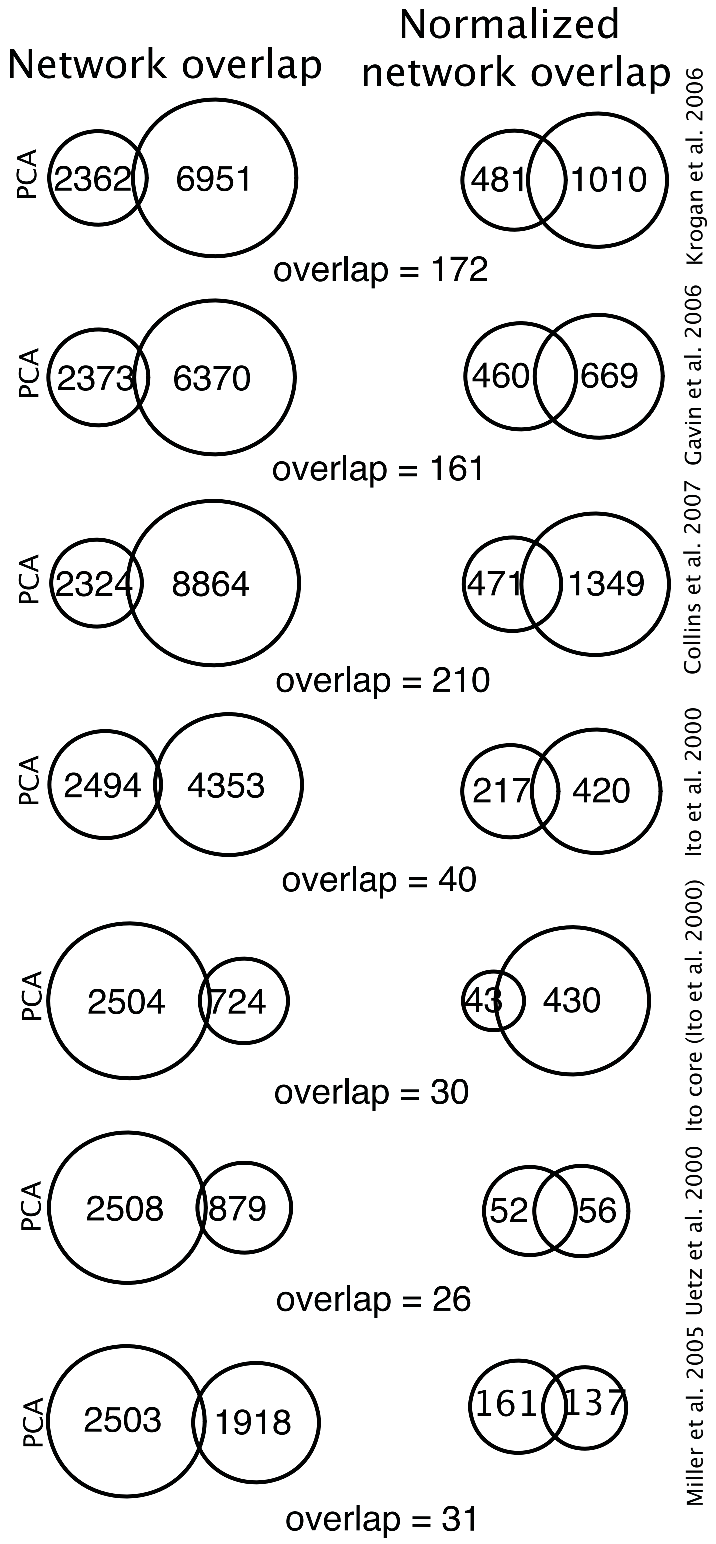
Supplementary figure 5: Quality assessment for the PCA networks and other PPI networks. The curve represents the total number of true positive interactions and the total number of false positive interactions as a function of the score thresholds for defining PPIs in the DHFR PCA screen (ROC curve). Values for published datasets are shown as well as values of the final DHFR PCA networks. Sources for the other networks are described in the Materials and Methods Section 13.



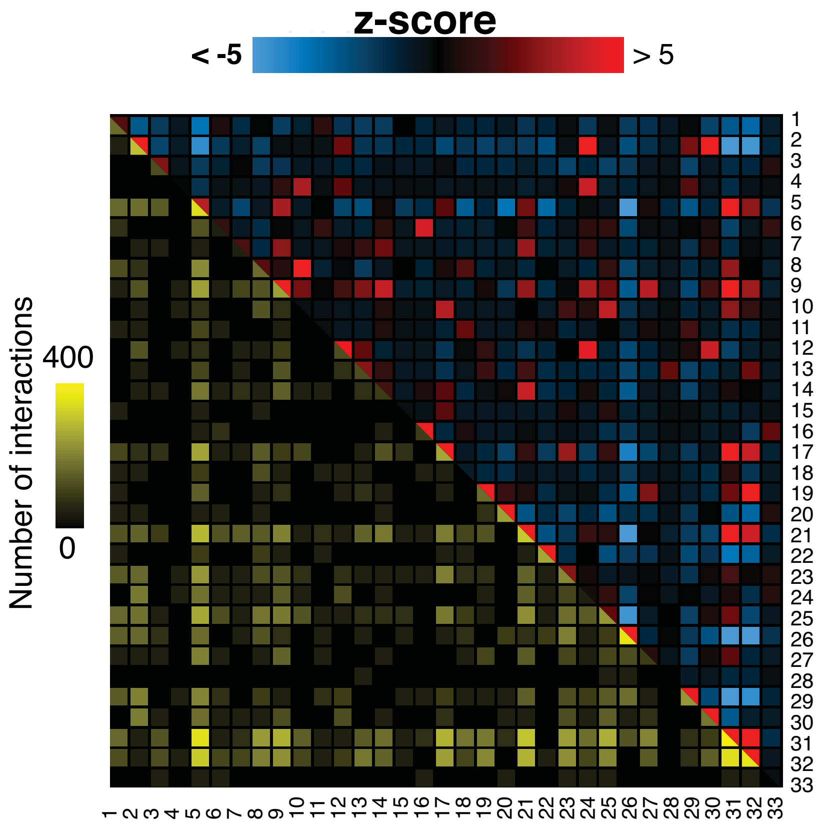
Supplementary figure 6. Distribution of protein abundance. The distribution of protein abundance for cells grown on the same (SC, SD + glucose (from (24))) medium used in the DHFR PCA screen of the entire proteome (black), proteins of the DHFR PCA network (blue) and proteins interaction with the control fragments (yellow).



Supplementary figure 7: Overlap of the DHFR PCA network with other large-scale experiments. (A) Most DHFR PCA PPIs are new, since they score 0 within the distribution of the number of times a known interaction has been independently deposited in major PPI repositories. Examples of interactions are shown above the bars. (B) The overlap of the DHFR PCA network is substantially increased when only the interactions that could be discovered are considered, i.e. only identified successful baits and preys are considered. Bars indicate the number of PPIs that could have been discovered by PCA. In red is the number of interactions that were discovered. Percentages indicate the percentage of interactions that were discovered by PCA out of the total possible.

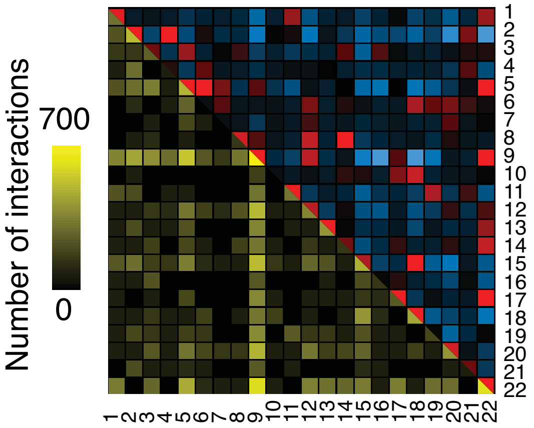


Supplementary figure 8: Overlap of the DHFR PCA network with other large-scale experiments. On the left is the overlap between the different networks. On the right are the same overlaps, but only for those interactions that could have been detected in both experiments; i.e. cases in which the interactions were tested for in both experiments.



Biological Process

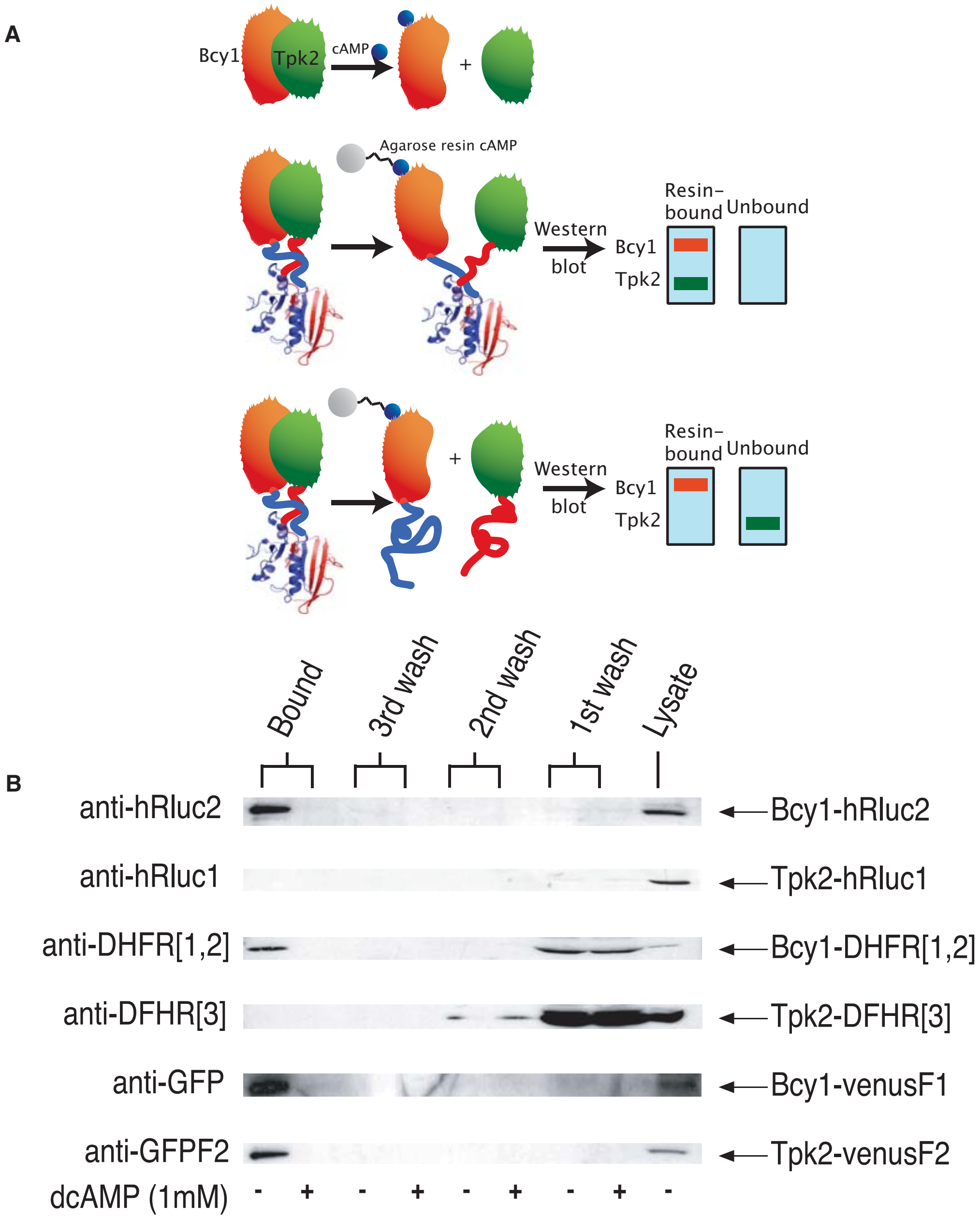
(1) DNA metabolism; (2) RNA metabolism; (3) amino acid and derivative metabolism; (4) anatomical structure morphogenesis; (5) unknown; (6) carbohydrate metabolism; (7) budding; (8) cell cycle; (9) cell wall organization and biogenesis; (10) cellular homeostasis; (11) cellular respiration; (12) conjugation; (13) cytokinesis; (14) cytoskeleton organization and biogenesis; (15) electron transport; (16) generation of precursor metabolites and energy; (17) lipid metabolic process; (18) meiosis; (19) membrane organization and biogenesis; (20) nuclear organization and biogenesis; (21) organelle organization and biogenesis; (22) protein catabolic process; (23) protein modification process; (24) pseudohyphal growth; (25) response to stress; (26) ribosome biogenesis and assembly; (27) signal transduction; (28) sporulation; (29) transcription; (30) translation; (31) transport; (32) vesicle-mediated transport; (33) vitamin metabolism;



Molecular function

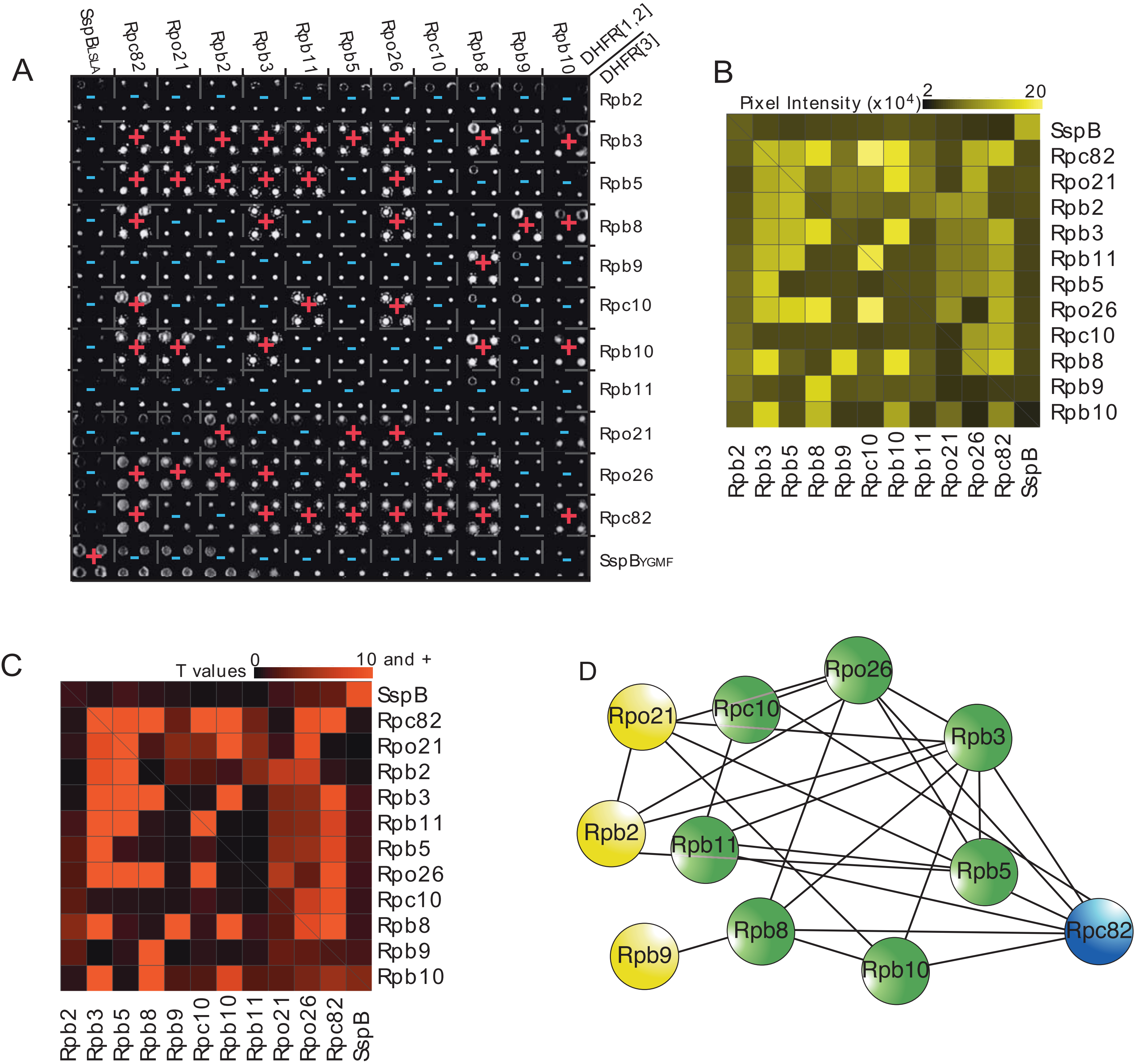
(1) DNA binding; (2) RNA binding; (3) enzyme regulator activity; (4) helicase activity; (5) hydrolase activity; (6) isomerase activity; (7) ligase activity; (8) lyase activity; (9) molecular_function; (10) motor activity; (11) nucleotidyltransferase activity; (12) oxidoreductase activity; (13) peptidase activity; (14) phosphoprotein phosphatase activity; (15) protein binding; (16) protein kinase activity; (17) signal transducer activity; (18) structural molecule activity; (19) transcription regulator activity; (20) transferase activity; (21) translation regulator activity; (22) transporter activity;

Supplementary figure 9: Interactions are enriched within Gene Ontology categories. The DHFR PCA network covers several classes of protein function, location and biological process. The colors above the diagonal represent positive and negative deviations from the expected number of interaction between two categories, Biological Process or Molecular function. A positive z-score indicates a larger number of interactions within or between two categories compared to a random network. A negative z-score indicates a smaller number of interactions than expected. A z-score of 2 or -2 corresponds to a P-value of 0.05 and a z-score of 5 or -5 to a P-value of 5×10-7. Values below -5 and above 5 were given these minimal and maximal values. z-scores were calculated by generating 10,000 random networks. Entries below the diagonal indicate the observed number of interactions on a log10 scale.



Tarassov et al. Supplementary figure 10

Supplementary figure 10: The DHFR PCA is reversible. (A) Upper panel. cAMP-mediated dissociation of the yeast PKA regulatory (Bcy1p) and catalytic (Tpk2p) subunits. Middle panel. Schematic representation of an irreversible PCA for PKA and predicted results of the Bcy1 subunit binding to cAMP-conjugated agarose beads. For the irreversible PCA, Bcy1 and Tpk2 dissociate but remain trapped by the folded PCA reporter protein. Lower panel. For a reversible PCA, reporter protein fragments unfold and dissociate when Bcy1 binds to cAMP-conjugated agarose beads and thus Bcy1 remains bound to the resin while Tpk2 is found in the unbound supernatant fraction. (B) The DHFR PCA is fully reversible. As reported previously (5), the Rluc PCA is reversible; Bcy1 is found in the cAMP-conjugated agarose fraction while Tpk2 is found in the wash. Precisely the same result is found for the DHFR PCA, suggesting that it is reversible, while Venus YFP PCA is irreversible.



Supplementary figure 11. RNA polymerase II complex reconstitution through DHFR PCA. (A) Results of the RNA polymerase II complex PCA network through an exhaustive screen for interactions among the ten subunits. Colonies for diploid strains that show resistance to methotrexate are indicated with red "+" and for those showing no resistance, with blue "-". (B) Mean colony pixel intensity values extracted from the high-resolution image in (A) by quantification of total colony pixel intensities. (C) t-scores for colony pixel intensities ranging from 0 to 10 and higher ($P < 0.0001$) resulting from the comparisons with control colonies. (D) Summary of the results of the RNA Pol II DHFR PCA screen where edges represent a physical interactions corresponding to a t-score of 4 and higher ($P < 0.05$) and nodes are the individual RNA Pol II subunits. Nodes colored in yellow, blue and green are respectively RNA Pol II exclusive proteins, RNA Pol III exclusive protein and RNA Pol II and RNA Pol III shared proteins.