**Supplementary Information**

**Table of Content:**
Figure S1  
Figure S2  
Figure S3  
Figure S4  
Figure S5  
Legend Table S1  
Legend Table S2  
Legend Table S3  
Legend Table S4  
Legend Table S5  
Legend Table S6  
Legend Table S7  
Supplementary materials and methods  
References to supplementary information  

**Figures and Tables**

**Figure S1. Phosphopeptide chip.** The structure of the phosphorylated peptide chip used to profile the substrate specificity of PTP trapping domain. A) Three identical sub-arrays are printed on the chip surface. B) Each sub-array is composed of a 4x4 grid of 400 spots. C) Each small grid unit contains the peptides and the control spots reported in the legend.
Figure S2. PLC-γ1, Gab1 and SHP2 are dephosphorylated by PTP1B in Src activated cell lines. A, left panel) HEK 293 cells were transiently transfected with a plasmid expressing the constitutive active Src-Y527F, (lanes 2 and 3) to increase tyrosine phosphorylation and co-transfected either with FLAG-tagged PTP1B (lanes 3) or with a corresponding amount of empty FLAG vector (lanes 1, 2). The input extracts were analyzed by SDS-PAGE and probed with antibodies against PLC-γ1 (WB: anti-PLC-γ1), Gab1 (WB: anti-Gab1), SHP2 (WB: anti-SHP2), Grb2 (WB: anti-Grb2), PTP1B (WB: anti-PTP1B), Src (WB: anti-Src) and anti-phosphotyrosine (WB: anti-4G10). A, right panel) 1 mg of protein extract was immunoprecipitated with anti-phosphotyrosine antibody 4G10 (IP: anti-4G10). The immunoprecipitated proteins were separated on SDS-PAGE, transferred onto nitrocellulose membranes and probed with antibodies against PLC-γ1 (WB: anti-PLC-γ1), Gab1 (WB: anti-Gab1), SHP2 (WB: anti-SHP2) and Grb2 (WB: anti-Grb2). Grb2 is used as a non-substrate control. The ratio between the intensity of the bands obtained with any specific antibody and with the anti-Grb2 antibody is reported as a bar diagram. B) PLC-γ1, Gab1, SHP2 and Grb2 proteins were immunoprecipitated from 1 mg of the cell lysates. The immunoprecipitated proteins were analyzed by western blot first with anti-4G10 antibody to detect the phosphorylation levels of each protein and then with protein specific antibodies. The ratio of phosphorylated to total protein is reported in a bar diagram.
Figura S3. 293T cells were co-transfected with a plasmid directing the synthesis of constitutively active Src or a plasmid control and with a mixture of two shRNA plasmids targeting the pTP1B gene (Mission Human library, SIGMA). Cell extracts were immunoprecipitated with antibodies against the four target proteins and revealed with an anti phosphotyrosine antibody (4G10). The band intensities were quantified and represented in a bar graph.
Figure S4. Stable cell lines expressing siRNA against PTP1B. HEK 293 cells were stably transfected with a plasmid expressing a short hairpin RNA targeting PTP1B. As a negative control we used a plasmid expressing a short hairpin RNA containing 5 bases that do not target any human gene. We selected 14 PTP1B interference clones and the corresponding protein extracts were immunoblotted with anti-PTP1B (A), and with anti-beta-tubulin (B) antibodies. The bar diagram depicts the decrease in PTP1B expression by arbitrarily setting to 100 the amount of PTP1B in the control lane. Among the fourteen clones that were analyzed we selected clone 11 showing the lowest PTP1B expression level with a down regulation of 96% when compared to a control cell line expressing a scrambled siRNA.

Figure S5. TC-PTP expression levels do not affect the SHP2-Gab1 complex. HEK 293 cells stably transfected with FLAG tagged TC-PTP (A) were treated for 24 hours with doxycycline to induce the expression of the phosphatase (lanes 3). HEK 293 cells were then induced five minutes with EGF 100 nanograms/ml (lanes 2 and 3). 1 mg of the protein extract of each sample was immuno-precipitated with Gab1 antibody (IP: anti-Gab1). The input and the Gab1-bound proteins (IP material) were immunoblotted with antibodies against Gab1 (WB: anti-Gab1) to verify IP and anti-SHP2 (WB: anti-SHP2) and 4G10 to monitor the level of phosphorylated Gab1. The expression of TC-PTP phosphatases was revealed using FLAG antibody (A). (B) The amount of immuno-precipitated Gab1 was used as a reference to quantify the SHP2 fraction bound to phosphorylated Gab1 in TC-PTP overexpression conditions (bar diagram in B).
Table S1. Phosphopeptides.
The 6400 phosphopeptides synthesized and arrayed on the chip. In row B, the sequence; in row C, the identification of the protein, in row D, the ELM reference; in row E, the Psite reference; in row F, the neural network prediction.

Table S2. Position Specific Scoring Matrix.
The PSSM used to score the phosphorylated peptide is reported as a 20 (aminoacid; first row) x 13 (peptide position; first column) matrix.

Table S3. PSSM score.
The 9530 human proteins (SPROT_ID) that are either annotated as phosphorylated proteins in the Phosphosite or PhosphoELM database or are nodes of the WINT_homo interactome were classified in ten groups according to their PSSM score. The score ranges (RANGE_SCORE) and the corresponding classes (CLASS) are indicated.

Table S4. Literature derived PTP1B interaction network.
The table reports the name (partner_name) and the Swissprot ID (partner_sprot) of the PTP1B partners described in the literature. The third column (PMID) contains the NCBI PMID of the articles reporting the supporting evidence. Finally, partners are classified as “interactor” if a physical association with PTP1B is demonstrated without any associated evidence of dephosphorylation and as “substrate” if dephosphorylation has been experimentally demonstrated.

Table S5. Weighted Interactome Distance (WID) score.
The 9530 human proteins (SPROT_ID) that are either annotated as phosphorylated proteins in the Phosphosite or PhosphoELM database or are nodes of the WINT_homo interactome were classified in ten groups (WID_CLASS) according to their network distance from PTP1B. The WID distance ranges (WID_DISTANCE) and the corresponding classes (CLASS) are indicated.

Table S6. Bayesian score.
The human proteins (identified with their SPROT_ID in the first column) were ranked according to the probability of being PTP1B substrates (BAYES_SCORE). Proteins that have been identified in the literature as PTP1B substrates are labeled with 1 in the third column (CLASS).

Table S7. Substrates.
List of predicted PTP1B substrates that are known to be involved in the EGFR and in the Insulin signaling pathways as annotated in the Reactome database. Already described PTP1B substrates, as reported in previously published work, are highlighted in bold. In red are the substrates that have been validated in this work.

High ranking predicted PTP1B substrates that are annotated by reactome in the EGFR and/or insulin pathways.
| Rank | UniProt AC | Gene Name | Bayes |
|------|------------|-----------|-------|
| 2    | P19174     | PLCG1     | 0.87608 |
| 22   | Q13480     | GAB1      | 0.85852 |
| 33   | P22681     | CBL       | 0.85852 |
| 44   | O14964     | HGS       | 0.85312 |
| 54   | P00533     | EGFR      | **0.83242** |
| 56   | P12931     | SRC       | **0.82817** |
| 66   | P27986     | PIK3R1    | 0.81842 |
| 74   | Q06124     | PTPN11    | 0.79219 |
| 84   | P01112     | HRAS      | 0.79011 |
| 87   | P42336     | PIK3CA    | 0.78946 |
| 95   | P04049     | RAF1      | 0.78937 |
| 110  | P29353     | SHC1      | 0.75998 |
| 112  | Q07889     | SOS1      | 0.75369 |
| 125  | P31946     | YWHAB     | 0.72951 |
| 176  | O75886     | STAM2     | 0.6254 |
| 177  | P60953     | CDC42     | 0.62095 |

**Supplementary materials and methods**

**Reagents, antibodies and plasmid constructs.**

Human insulin solution was from SIGMA. Polyclonal antibodies for SHP2 (1:2500 dilution in western blots), Grb2 (WB 1:5000) and PTP1B (WB 1:2500) were purchased from BD Transduction Laboratories; polyclonal antibodies for Gab1 (WB 1:1000) and PLC-\(\gamma\)1 (WB 1:1000) were purchased from Cell Signaling; polyclonal antibody against beta tubulin (WB 1:500) was purchased from Santa Cruz Biotechnology; monoclonal anti-phosphotyrosine antibody (4G10) (WB 1:1000) were purchased from Upstate Biotechnology. The anti-rabbit (WB 1:5000) and anti-mouse (WB 1:2500) secondary antibodies were purchased from Jackson Immunoresearch. The goat anti-rabbit GST-horseradish peroxidase conjugate was purchased from SIGMA. Antibodies against Gab1, PLC-\(\gamma\)1 and 4G10 were diluted in a blocking solution containing 3% BSA in PBS, while the other antibodies were diluted into 5% non-fat milk dissolved in PBS. To generate a vector expressing fusion between PTP1B and the 3XFlag epitope, we amplified the PTP1B cDNA using the following primers: 5’ CTAGCAGCGGCGGCGGCCGGTCATGGAGATGGAA 3’ inserting the sequence recognized by NolI restriction enzyme in front of the phosphatase translation initiation site, and 5’ TGCTAGTCTAGATGTGTTGCTGTGGACAGGAACCT 3’ inserting the sequence recognized by the XbaI restriction enzyme. The amplified fragment was inserted into a p3XFLAG-CMV™-14 expression vector from Sigma Aldrich linearized by cleavage with the enzymes NolI and XbaI; the cDNA encoding the Y527F Src mutant is cloned in pSGT (1). Protein A-Sepharose beads used for immunoprecipitations were obtained from Sigma. To generate the vector for inducible expression of
PTP1B in the Flp-In plasmid (Invitrogen), we amplified the PTP1B sequence from the p3XFLAG vector using the following primers: CTAGCGGATCCGCCTATCGGACATGGA and TGCTAGGCAGGCGCTACTTTATCGCTGCTACCTTGTAATCTGTGTTGCTGTGCTGTGAACAGGAAC. The PTP1B inhibitor was 3-(3,5-Dibromo-4-hydroxy-benzoyl)-2-ethyl-benzofuran-6-sulfonicacid-(4-(thiazol-2-ylsulfamyl)-phenyl)-amide, used at 250 microM, 1 hr before cell lysis.

**GST fusion protein purification.**

The pGEX-4TK vector encoding the mutated catalytic domain of PTP1B was provided by Rob Hooft van Huijsduijnen. The pGEX-4TK vector encoding the SH2 domains of SHP2 was provided by Bruce Mayer. Escherichia coli BL21 Rosetta were transformed, and a single colony was grown in 25 ml of LB at 37 °C until an OD of 0.5. Protein production was performed at 30 °C for 4 hours after the addition of IPTG at a final concentration of 250 µM. The bacteria were pelleted and resuspended in lysis buffer (50 mM Tris, pH 8.0, 5 mM EDTA, 0.1% Triton X-100, 150 mM NaCl) supplemented with a proteinase inhibitor mixture (Roche Applied Science) and lysed by lysozyme treatment (200 µg/ml) for 1 hour on ice followed by three rounds of sonication. The lysate supernatant was incubated for more than 2 hours with 100 µl of a 50% solution of glutathione-Sepharose beads (Amersham Biosciences) at 4°C. Finally, beads were extensively washed in PBST (PBS, 0.1% Triton X-100), and the GST fused PTPs were eluted in 50 mM Tris, pH 8.0, with 10 mM glutathione. Glycerol was added at a final concentration of 20%; the amount of proteins produced was determined with a Bio-Rad protein assay, and aliquots were stocked at −80 °C until use.

**Generation of a stable cell line expressing a siRNA targeting PTP1B mRNA.**

To generate a stable cell line knocked down for PTP1B expression, HEK 293 cells were transiently transfected with the PLKO.1 plasmid encoding a siRNA targeting the PTP1B transcript (MISSION RNAi SIGMA, PTP1B siRNA CCGGCTAACCACATGCGGTACTTTCTCGAGAAAGTGACCGCATGTGTTAGGTTTTT; scrambled control CCGGCAACAAGATGAAGAGCACCAACTCGAGTTGGTGCTCTTCTGTTGTTGTTTTTT). Positive clones were selected following the manufacturer’s instructions (SIGMA). Western blot monitored the PTP1B levels in 14 isolated transfectants and the clone with the lowest PTP1B expression level was selected for the experiments described here. The experiment reported in Figure S3 was carried out with a mixture of two siRNA with the following sequence: CCGGGGCTGCTTGCTATATGCTACTCGAGTAAGGCACTATAGCAGAGCAGCTTTTT CCGGGAAGCCCAAAGGAGTTACATTCTCGAGAATGTAACTCCTTTTGCTCTTTTT.

**Cell culture, transfection and extract analysis.**

HEK (Human Embryonic Kidney) 293 cells were purchased from ATCC and cultured in a humidified atmosphere (5% CO2) at 37°C in HEK-293 Dulbecco's Modified Eagle Medium (Gibco/Invitrogen), supplemented with 10% fetal bovine serum (Sigma Aldrich) and 0.1% penicillin/streptomycin (Invitrogen). HEK 293 cells were transfected by the calcium phosphate method as previously described. To study the effect of PTP1B upon insulin stimulation, confluent and transfected HEK 293 cells were incubated for 20 hours in serum-free medium and then treated with 100 ng/ml of insulin at appropriate times. Cells were then washed with ice-cold phosphate-buffered saline (PBS) and lysed in RIPA buffer (150 mM NaCl, 50 mM Tris-HCl, 1% NP-40, 0.25% Na deoxyxlate) supplemented with 2mM sodium vanadate, 1 mM NaF, Protease inhibitor cocktail 200x (Sigma), Phosphatases inhibitor Cocktail I and II 100x (Sigma). The samples were kept in ice and then centrifuged at
15000 rpm at 4°C for 30'. The supernatant was collected and the total amount of protein determined by the Bradford colorimetric assay (Bio-Rad). For immunoprecipitation experiments, the total protein extract was pre-cleaned by incubation for 2 hours at 4°C with Protein A-Agarose beads (Sigma) pre-equilibrated with lysis buffer, then recovered by centrifugation at 5000 g for 10 minutes at 4°C and the antibody was then added in order to immuno-precipitate the protein of interest overnight at 4°C. After centrifugation for 3 minutes at 4000 g at 4°C, the supernatant was removed and the beads washed four times for 3 minutes with ice cold lysis buffer. For immunoblotting, protein samples were denatured by adding 3x SDS sample buffer with 2-mercaptoethanol and boiling for 10 minutes. The beads were removed by centrifugation at 15000 for 2’. The denatured material was subjected to SDS-PAGE, then the gel transferred to a nitrocellulose membrane. Membranes were incubated in blocking solution (5% non-fat milk in PBS or 3% BSA in PBS) for 1 hour at room temperature, and then incubated with primary antibodies overnight. After three washes with PBS/0.05% Tween20, the membranes were incubated with horseradish peroxidase-conjugated secondary antibody for 1 hour at room temperature and visualized by enhanced chemiluminescence (Amersham Pharmacia Biotech) on LAS-3000 imaging system (Fujifilm Life Science).

For far western assays, the nitrocellulose membranes blocked as for immunoblotting and then were probed with purified GST fused SH2 domains for 1 hour at room temperature. After three washes with PBS containing 0.05% Tween20, the membranes were incubated with the primary anti GST antibody for 1 hour and washed three times. After incubation with horseradish peroxidase-conjugated secondary antibody for 1 hour at room temperature, bands were visualized by enhanced chemiluminescence (Amersham Pharmacia Biotech) on a LAS-3000 imaging system (Fujifilm Life Science).

For the pull-down assays confluent HEK 293 cells were incubated for 20 hours in serum-free medium and then treated with 100 ng/ml insulin at 5 min. Cells were then washed with ice-cold phosphate-buffered saline (PBS) and lysed in RIPA buffer (150 mM NaCl, 50 mM Tris-HCl, 1% NP-40, 0.25% Na deoxycholate) supplemented with 2mM sodium vanadate, 1 mM NaF, Protease inhibitor cocktail 200X (Sigma), inhibitor Phosphatases Cocktail I and II 100X (Sigma). The samples were kept on ice for 30’, then centrifuged at 15000 rpm, 4°C, for 30’. The supernatant was collected and the total amount of protein determined by the Bradford colorimetric assay (Bio-Rad). The whole cell extract was pre-cleared on GST bound to glutathione-Sepharose 4B beads with rocking at 4°C for 2 hours, then the supernatant was collected by centrifugation for 5’ at 4000 g at 4°C. Each GST-fusion protein bound to glutathione-Sepharose 4B beads was blocked by incubating with 3% BSA with rocking at 4°C for 2 hours, then after centrifugation for 3’ at 4000 g, 4°C, the dry beads were equilibrated with lysis buffer. Equal parts of cell extract were added to 100 µg of each fusion protein with rocking at 4°C for 12 hours. The supernatant was discarded by centrifugation and the beads were washed 3 times with 1.5 ml of lysis buffer for 3’ at 4000 g, 4°C, then the dry beads were resuspended in 50 µl of 3X SDS sample buffer, boiled for 10’, and the eluted proteins collected by centrifugation at 15000 rpm for 10’.

**Generation of stable inducible HEK293 T cell line expressing controlled levels of PTP1B and TC-PTP using the Flp-In T Rex system (Invitrogen).**

The Flp-In T-REx system (FIT) (Invitrogen) was used for stable, inducible PTP1B or TC-PTP expression in HEK 293 cells. PTP1B or TC-PTP cDNAs were inserted into the polylinker of plasmid pcDNA5/FRT/TO (Invitrogen). The stably transfected cell lines were generated based on the Flp-In T-REx-293 cell line according to the FIT system protocols (Invitrogen). To turn on phosphatase expression, cells were cultivated for at least 20 h in regular growth medium supplemented with 1 µg/ml doxycycline.
Assembly of WINT: a weighted human interactome.

A weighted human interactome was assembled by combining the interactions imported from four major protein-protein interaction (PPI) databases: DIP (2), HPRD (3), IntAct (4) and MINT (5). Three of these repositories are part of the IMEx consortium and have agreed on common guidelines for curating protein-protein interactions (6). These guidelines recommend that each record be accompanied by a complete description of the experiments reporting the evidence for the interaction, including the interaction detection method and the interaction type. All this information is valuable for assessing data reliability and the quality of the supporting evidence. Moreover, IMEx partners use common controlled vocabularies and release data in a common format (Protein Standard Initiative-Molecular Interaction (PSI-MI) format); as such, the records in these datasets can be integrated smoothly (6). Conversely, the HPRD database releases a larger number of records of interactions in humans, but does not follow the described guidelines and each interaction is associated to an unmapped list of methods and PubMed references. We have included the HPRD database but, in those cases where it was impossible to map a method to a single PubMed reference (e.g., more than one method associated to more than one PubMed reference), we conservatively assigned the basic method in the PSI-MI controlled vocabulary "experimental detection method" to the interaction. To solve the problem of overlap between database entries (i.e., publications curated by more than one database), we imported the entries in the following order, excluding publications imported in previous steps: MINT (October 2008 release), IntAct (psi2.5 XML release of 22/08/2008), DIP (MITAB files: Hsapi20080407.txt, Mmusc20080407.txt), and HPRD (release 090107). In a second step the interactions demonstrated in model organisms have been mapped to the human orthologs by using the procedure implemented in HomoMINT (7). For orthology mapping we retrieved the homology tables via the BioMArt Ensembl webservice (datasets: C. elegans, mouse, rat, yeast, drosophila, X laevis, cow, dog, chicken). The percent of identity between a human protein and its ortholog in the model organism is kept in the HomoMINT orthology table to be used later for the reliability scoring algorithm. To rank the interaction pairs we have recently introduced a composite score. Each interaction is assigned a score ranging from 0 to 1 by taking into account the “combined experimental evidence”, as described at: http://mint.bio.uniroma2.it/mint/doc/MINT-confidence-score.html (8).

The weight associated to any interaction in the interactome is an estimate of the interaction reliability. Thus, a pair of interacting proteins linked by a high-weight interaction may be thought to be more tightly connected or closer to each other than a pair of proteins connected by a low-weight interaction. Following this analogy, we defined the distance between two nodes connected by an edge of weight \( w \) as: \( d = 1 - w \). Since weights ranged from 0 to 1, also distances fell within the [0, 1] interval. We extended this notion of distance to any pair of nodes A and B for which at least one path going from A to B existed in the graph: the total distance was computed with the Dijkstra's algorithm (9), by adding up the distances associated to all the edges forming the path. When multiple alternative paths connecting A and B could be found, the minimum total distance was chosen. In this way, we could assign a weighted interactome distance (WID) from PTP1B to each protein in the human interactome that could be connected to the phosphatase with a path composed of a finite number of edges (7522 out of 7866 proteins) (see supplemental Table 4).

Training the Bayesian model.

We have built a Bayesian model capable of predicting whether a protein is a likely PTP1B substrate in vivo, based on two features: the score assigned by the PSSM encoding the experimentally derived consensus to the short amino acid sequences (13-mer) flanking any known phosphorylated tyrosine by
the PSSM encoding the experimentally derived consensus, and the distance (WID) separating the candidate substrate protein and the phosphatase in the context of the WINT human interactome.

PSSM and WID scores can be regarded as heterogeneous and independent pieces of evidence on the basis of which the likelihood for a certain protein to be an enzymatic target of PTP1B may be assessed. Since the two features are indeed statistically independent, we thought it appropriate to integrate them in a single Bayesian probabilistic model. To this aim, we collected from the literature, with the aid of simple text mining software, a set of 34 proteins, which are known to be dephosphorylated in vivo by the PTP1B phosphatase. A complete training set was generated by adding to this positive high-confidence set of known PTP1B substrates a negative set of 340 proteins, randomly selected from the approximately 9,500 which could be assigned either a PSSM or a WID score. It is reasonable to assume that most of the proteins in the negative set are not PTP1B substrates, given the presumably very low frequency of PTP1B substrates in the human proteome. To minimize the variability due to the random sampling of negative instances, we repeated the selection of negative examples 10 times and thus generated 19 different training sets. Before training, each feature was discretized in 10 equal frequency bins. Missing values were labeled 'missing' and put in an ad-hoc bin (see supplementary Tables S3 and S5). We trained a Bayesian classifier on each training set and evaluated their performance by 10-fold cross-validation (10 cross-validation runs for each training set, results were averaged), obtaining an average Area Under the Curve (AUC) of 0.91. Finally, we predicted on the entire protein set (9,530 proteins) with each of the 10 trained Bayesian models and averaged the results to obtain a single final score for each protein in the set (Supplementary Tables S6).

A binary Bayesian classifier produces two outputs for each classified instance: the posterior probability of the instance being positive (i.e., in our case, the probability of being a substrate of the PTP1B phosphatase), given the evidence, and the posterior probability of the instance being negative, given the evidence. Since these are probabilities of mutually exclusive and collectively exhaustive events (any protein is necessarily either a substrate of PTP1B or not, no further option exists), they sum up to one: thus, the classifier assigns a class label (“positive” or “negative”) to an instance simply by looking at which of the two aforementioned posterior probabilities is greater than 0.5. However, the actual value of the posterior probability is dependent upon the prior probability, according to

\[ P(C|E) = P(E|C) \times P(C) \]

where \( P(C|E) \) is the posterior probability of belonging to class C, given the evidence E, \( P(E|C) \) is the probability of observing the evidence E, given that the instance belongs to class C, and \( P(C) \) is the prior probability of belonging to class C. In our case, the prior probability of being an enzymatic target of the PTP1B phosphatase was unknown and we arbitrarily set it to 0.1. Thus, it would have made little sense if we selected the predicted substrates by looking at the actual posterior probability values. Since the mutual ordering of classified instances is not affected by the particular choice of prior probability, we ordered the proteins according to their posterior probability of being positive and then we decided to classify as predicted substrates those proteins whose posterior probability was above a certain threshold. We set the threshold so as to assure that at least half of the golden standard PTP1B substrates would be correctly classified as positive (50% sensitivity threshold). At the so chosen threshold of 0.54, we obtained a list of 236 putative PTP1B enzymatic targets, with a specificity of 98% (see supplemental Table S6).
Literature curation.

We have developed a simple text mining approach to search for information related to the PTP1B phosphatase. PubMed database was queried with the “phosphatase” string and with the synonyms of the protein PTP1B retrieved from PIR (10), ProThesaurus (11) and GENIA (12) databases. The procedure retrieved ~ 147000 results that were combined and downloaded in XML format. A Python script was designed to automatically parse the title and abstract fields of the XML formatted PubMed entry. The script was organized in 4 independent modules.

1) Protein names recognition: it searches for words in the manuscript abstract matching a list of strings containing the names associated to the UniProt entry of PTP1B. Capitalized characters, hyphens, parentheses were ignored in the string matching step.

2) Tyrosine phosphorylation: This module is designed to capture evidence that the article is describing experiments about a specific tyrosine residue. Regular expression patterns were compiled to search for the strings “tyrosine”, “pY”, “tyr”, followed by a number. Capitalized characters, hyphens, parentheses were ignored in the string matching step.

3) Interaction terms: We compiled a list of verbs or nouns describing an interaction. This is largely based on the list from Temkin et al. (13). Some terms were removed from the original set as they were not suitable for our own definition of an interaction (e.g., ‘cleaves’, ‘expressed’, ‘severed’).

4) Interaction methods: The terms related to methods commonly used to identify protein interaction was compiled from the PSI-MI Vocabularies. Additional strings (e.g., ‘co-precipitation’, ‘X-ray’) were added during the human review process upon observation of interaction phrases.

Each matched evidence was organized in a tab-delimited file with a score reflecting the number of positive hits in each parsing module. The abstracts with no hits were discarded while those with at least one positive hit were reviewed by an expert curator to discard false positive matches. The 84 selected manuscripts were curated according to the PSI-MI standards (4) and the interactions reported therein were stored in the MINT database (see supplemental Table S4) (5).

References to supplementary information

1. Gonfloni, S., Williams, J. C., Hattula, K., Weijland, A., Wierenga, R. K., and Superti-Furga, G. (1997) *Embo J* 16(24), 7261-7271
2. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004) *Nucleic Acids Res* 32(Database issue), D449-451
3. Prasad, T. S., Kandasamy, K., and Pandey, A. (2009) *Methods Mol Biol* 577, 67-79
4. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., and Hermjakob, H. (2007) *Nucleic Acids Res* 35(Database issue), D561-565
5. Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007) *Nucleic Acids Res* 35(Database issue), D572-574
6. Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-Aryamontri, A., Salwinski, L., Nerothin, J., and Hermjakob, H. (2007) *Proteomics* 7 Suppl 1, 28-34
7. Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., and Cesareni, G. (2005) *BMC Bioinformatics* 6 Suppl 4, S21
8. Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2009) *Nucleic Acids Res*
9. Dijkstra, E. W. (1959) *Numerische Mathematik* **1**(1), 269-271
10. Barker, W. C., Garavelli, J. S., Hou, Z., Huang, H., Ledley, R. S., McGarvey, P. B., Mewes, H. W., Orcutt, B. C., Pfeiffer, F., Tsugita, A., Vinayaka, C. R., Xiao, C., Yeh, L. S., and Wu, C. (2001) *Nucleic Acids Res* **29**(1), 29-32
11. Fundel, K., and Zimmer, R. (2006) *BMC Bioinformatics* **7**, 372
12. Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003) *Bioinformatics* **19 Suppl 1**, i180-182
13. Temkin, J. M., and Gilder, M. R. (2003) *Bioinformatics* **19**(16), 2046-2053