

# Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets

T K B Gandhi<sup>1</sup>, Jun Zhong<sup>2</sup>, Suresh Mathivanan<sup>1</sup>, L Karthick<sup>1</sup>, K N Chandrika<sup>1</sup>, S Sujatha Mohan<sup>1</sup>, Salil Sharma<sup>1,2</sup>, Stefan Pinkert<sup>3</sup>, Shilpa Nagaraju<sup>1</sup>, Balamurugan Periaswamy<sup>1</sup>, Goparani Mishra<sup>1</sup>, Kannabiran Nandakumar<sup>1</sup>, Beiyi Shen<sup>2</sup>, Nandan Deshpande<sup>1</sup>, Rashmi Nayak<sup>1</sup>, Malabika Sarker<sup>1</sup>, Jef D Boeke<sup>5,6</sup>, Giovanni Parmigiani<sup>4</sup>, Jörg Schultz<sup>3</sup>, Joel S Bader<sup>5,7</sup> & Akhilesh Pandey<sup>2,4</sup>

**We present the first analysis of the human proteome with regard to interactions between proteins. We also compare the human interactome with the available interaction datasets from yeast (*Saccharomyces cerevisiae*), worm (*Caenorhabditis elegans*) and fly (*Drosophila melanogaster*). Of >70,000 binary interactions, only 42 were common to human, worm and fly, and only 16 were common to all four datasets. An additional 36 interactions were common to fly and worm but were not observed in humans, although a coimmunoprecipitation assay showed that 9 of the interactions do occur in humans. A re-examination of the connectivity of essential genes in yeast and humans indicated that the available data do not support the presumption that the number of interaction partners can accurately predict whether a gene is essential. Finally, we found that proteins encoded by genes mutated in inherited genetic disorders are likely to interact with proteins known to cause similar disorders, suggesting the existence of disease subnetworks. The human interaction map constructed from our analysis should facilitate an integrative systems biology approach to elucidating the cellular networks that contribute to health and disease states.**

Proteins generally function as part of larger protein complexes that are, in turn, part of a more extensive network. One of the important goals of functional genomics is to identify the complete protein interaction network, or interactome, of an organism. Large-scale genome-wide protein interaction studies carried out in yeast<sup>1–4</sup>, fly<sup>5–7</sup> and worm<sup>8</sup> have provided insights into the complexity of protein interaction networks. However, the yeast two-hybrid (Y2H) system, an important component

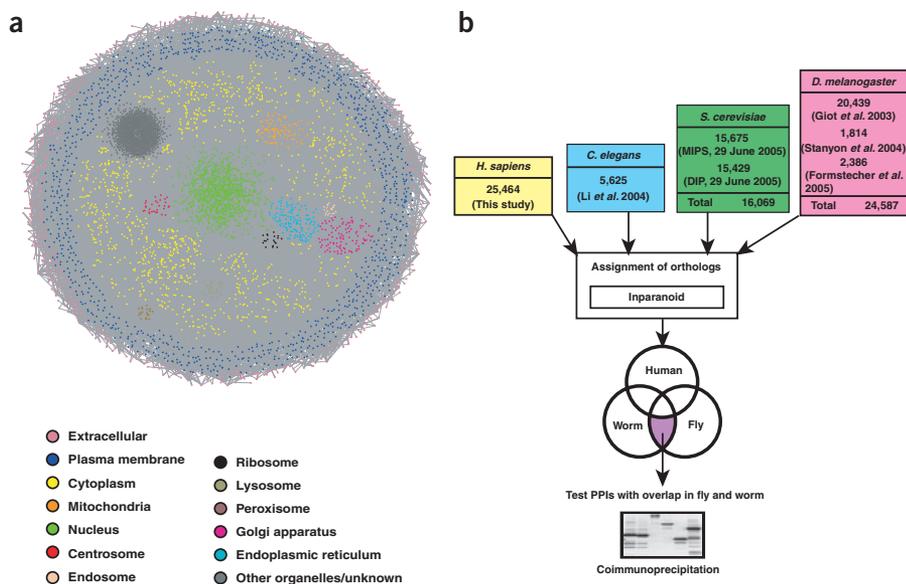
of high-throughput interaction studies, is subject to a high rate of false positives (roughly 50%)<sup>9</sup>. One way to evaluate the results of Y2H assays is to compare them to biochemically proven interactions. Unfortunately, very limited biochemical data exist for the worm and fly, as they are not amenable to such experiments. In humans, there exist a large number of published interactions accumulating from the study of individual proteins during the last few decades. We therefore used interaction data derived from the biomedical literature to develop a comprehensive map of protein interactions in humans.

## The human protein interaction network

We previously developed the Human Protein Reference Database (HPRD)<sup>10</sup> resource, which contains experimentally verified protein-protein interactions (PPIs) obtained from manual curation of the literature. This database contains 24,385 binary human PPIs, derived from Y2H assays, *in vitro* experiments (such as glutathione *S*-transferase (GST) pulldown) and *in vivo* experiments (such as coimmunoprecipitation). More than 95% of the interactions in the HPRD are based on individual *in vivo* or *in vitro* experiments (the number of interactions derived solely from Y2H assays is <5%). We supplemented these data with nonredundant human PPIs from several other databases: Biomolecular Interaction Network Database (BIND; 4,644 interactions)<sup>11</sup>, Database of Interacting Proteins (DIP; 1,049 interactions)<sup>12</sup>, Munich Information Center for Protein Sequences (MIPS; 334 interactions)<sup>13</sup>, Molecular Interactions Database (MINT; 3,544 interactions)<sup>14</sup> and IntAct<sup>15</sup> (2,420 interactions). Many of the PPIs present in BIND (85%), DIP (73%), MIPS (91%), MINT (83%) and IntAct (90%) were already present in the HPRD (T.K.B.G., S.M., K. Kumaran, S. Suresh, R.N. and A.P., unpublished data). For PPIs present in these databases but not in the HPRD, we carried out a literature search to confirm the interactions before including them in the HPRD. We were able to confirm the large majority (>95%) of these interactions where PubMed identification numbers were provided. The combined dataset contained 26,908 PPIs, of which 25,464 interactions could be mapped to unique gene identifiers and were used in this analysis.

To visualize the human protein interaction network, we took advantage of the available information about the subcellular localization of the majority of these proteins (from the literature, as annotated in the HPRD). **Figure 1a** shows the human protein interaction network, which is the most detailed interactome described thus far. The longest distance,

<sup>1</sup>Institute of Bioinformatics, International Technology Park, Bangalore 560 066, India. <sup>2</sup>McKusick-Nathans Institute of Genetic Medicine and the Departments of Biological Chemistry and Pathology, Johns Hopkins University, Baltimore, Maryland 21205, USA. <sup>3</sup>Department of Bioinformatics, Biozentrum, University Würzburg, Am Hubland, 97074 Würzburg, Germany. <sup>4</sup>The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, <sup>5</sup>The High Throughput Biology Center, <sup>6</sup>Department of Molecular Biology and Genetics and <sup>7</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21205, USA. Correspondence should be addressed to A.P. (pandey@jhmi.edu).



**Figure 1** Human protein interactome. (a) Human protein interaction map representing the overall network of PPIs in humans. Nodes represent proteins colored according to primary site of subcellular localization; edges indicate PPIs. (b) Summary of yeast, worm, fly and human PPI datasets and strategy used to analyze interaction data.

or diameter, between any two random nodes in this network was 18; the characteristic path length, or average of the minimum distances between all nodes, was 5.14.

### Overlap of PPIs in yeast, worm, fly and human datasets

Physical interactions of protein pairs in one organism are expected to be conserved in other related organisms<sup>16</sup>. An interspecies comparison of PPI data from yeast, worm and fly was recently carried out to identify conserved subnetworks<sup>17</sup>. Because of the availability of a large human PPI dataset, we investigated the extent to which the human PPIs overlap with those reported in the yeast, worm and fly datasets. The yeast dataset consists of 16,069 interactions derived from literature-based curation and high-throughput experimental methods stored in DIP<sup>12</sup> and MIPS<sup>18</sup>; the worm dataset contains 5,625 interactions obtained from a single high-throughput Y2H study<sup>8</sup>; and the fly dataset contains 24,587 fly interactions from three high-throughput Y2H studies<sup>5–7</sup>. To identify overlaps across species, orthologs must first be accurately identified, which is itself a difficult task. To identify orthologs, we used the Inparanoid database<sup>19</sup>, which is based on an all-versus-all BLAST search following by clustering into orthologous groups. Overall, there were 1,131 yeast, 1,079 worm and 3,347 fly genes that have high-confidence human orthologs. The numbers of orthologous protein pairs between the human, worm, fly and yeast genomes are shown (Table 1). These orthologous pairs were used to systematically examine the

**Table 1** Overlap of orthologs across human, worm, fly and yeast genomes

Worm	1,079	—	—
Fly	3,347	2,882	—
Yeast	1,131	1,318	1,392
	Human	Worm	Fly

The table shows the number of pairwise high-confidence orthologs between the indicated species as assigned by the Inparanoid database<sup>19</sup>.

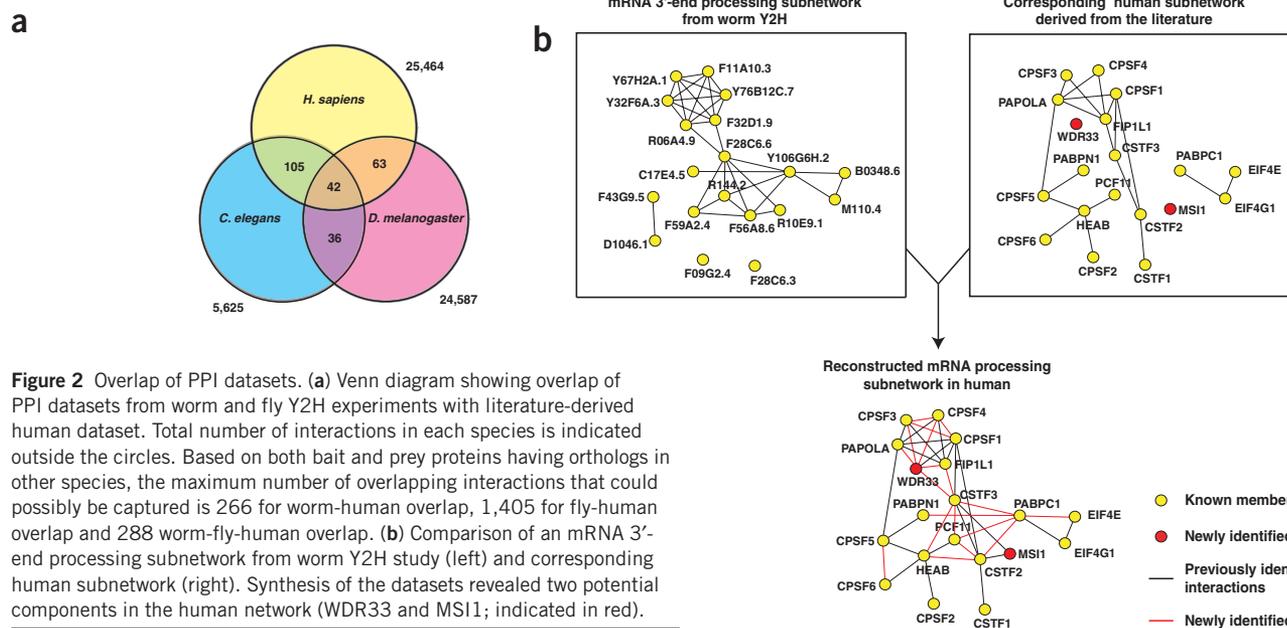
overlap among a total of >70,000 interactions, which included 25,464 human, 16,069 yeast, 5,625 worm and 24,587 fly pairwise interactions (Fig. 1b). For the Y2H studies, we initially included both high- and low-confidence interaction pairs when determining the overlap to avoid excluding any possible overlap.

As indicated by a Venn diagram showing the overlap between the human, fly and worm interaction datasets (Fig. 2a), 42 PPIs were common to the human, worm and fly datasets (Supplementary Table 1 online) and 16 were common to the human, worm, fly and yeast datasets. This relative lack of overlap is somewhat surprising given the large number of interactions in each of these datasets; there is a substantial number of genes with orthologs in the four species (Table 1). One of the possible reasons for the low overlap is that the baits used in the Y2H study did not have a human ortholog: only 888 of the 3,750 baits used to generate the fly interaction dataset and 424 of the 1,873 baits used to generate the worm interaction dataset have orthologs in the human interactome. We also determined the number of human PPIs for which both interacting proteins had orthologs in other species and were present as bait, prey or both in the Y2H datasets. This number indicates the maximum amount of overlap that could possibly be detected. Sixty-three of the maximum of 1,405 PPIs were detected as overlaps in the human-fly comparison, 105 of the maximum of 266 PPIs were detected as overlaps in the human-worm comparison and 42 of the maximum of 288 PPIs overlapped in all three species (Fig. 2a). When we reanalyzed the data using only the high-confidence interactions, we found that the overlap dropped considerably in every case. For example, the overlap between worm and human alone dropped from 105 to 15, the overlap between fly and human alone dropped from 63 to 43 and the overlap between worm, fly and human dropped from 42 to 12.

Cataloging of the clusters that overlapped between the human and yeast, worm or fly datasets revealed that most of the subnetworks are involved in DNA replication and repair, transcription, mRNA processing and protein folding or transport (Supplementary Figs. 1–3 online). Most of the interactions that overlap across species are involved in a limited number of biological processes, a fact that is likely to reflect the biased nature and limitations of the Y2H analysis and has been noted by others<sup>20–22</sup>.

### Prediction of human PPIs from worm and fly datasets

To determine whether individual subnetworks could provide information on related protein complexes in humans, we looked for highly connected subnetworks in the worm and fly datasets. In the worm Y2H dataset, we found a densely interconnected complex that contained several components of the pre-mRNA 3'-end processing machinery (Fig. 2b, left). The corresponding literature-derived human complex was more sparsely connected (Fig. 2b, right). Two worm proteins, R06A4.9 and R10E9.1, are part of this complex, but their human orthologs, WDR33 and MS11, had not previously been shown to participate in this process. Because R06A4.9 is connected to six other components of this complex, its human ortholog, WDR33, is likely to be a *bona*



*vide* component of the human complex. The other worm protein, R10E9.1, contains RNA recognition motifs and interacts with two other proteins in the worm complex. This suggests that MS1 has a role in RNA recognition in the human 3'-end processing complex. In addition, Y106G6H.2, the worm ortholog of the poly(A)-binding protein PABPC1, is connected to several proteins in the worm subnetwork, again indicating that PABPC1 could be a member of this complex in humans. These data allowed us to build a consensus of the complex involved in 3'-end mRNA processing in humans (Fig. 2b, bottom). Highly interconnected clusters such as these could be extrapolated to other species in which the interaction data is sparser and subsequently tested by experimental methods.

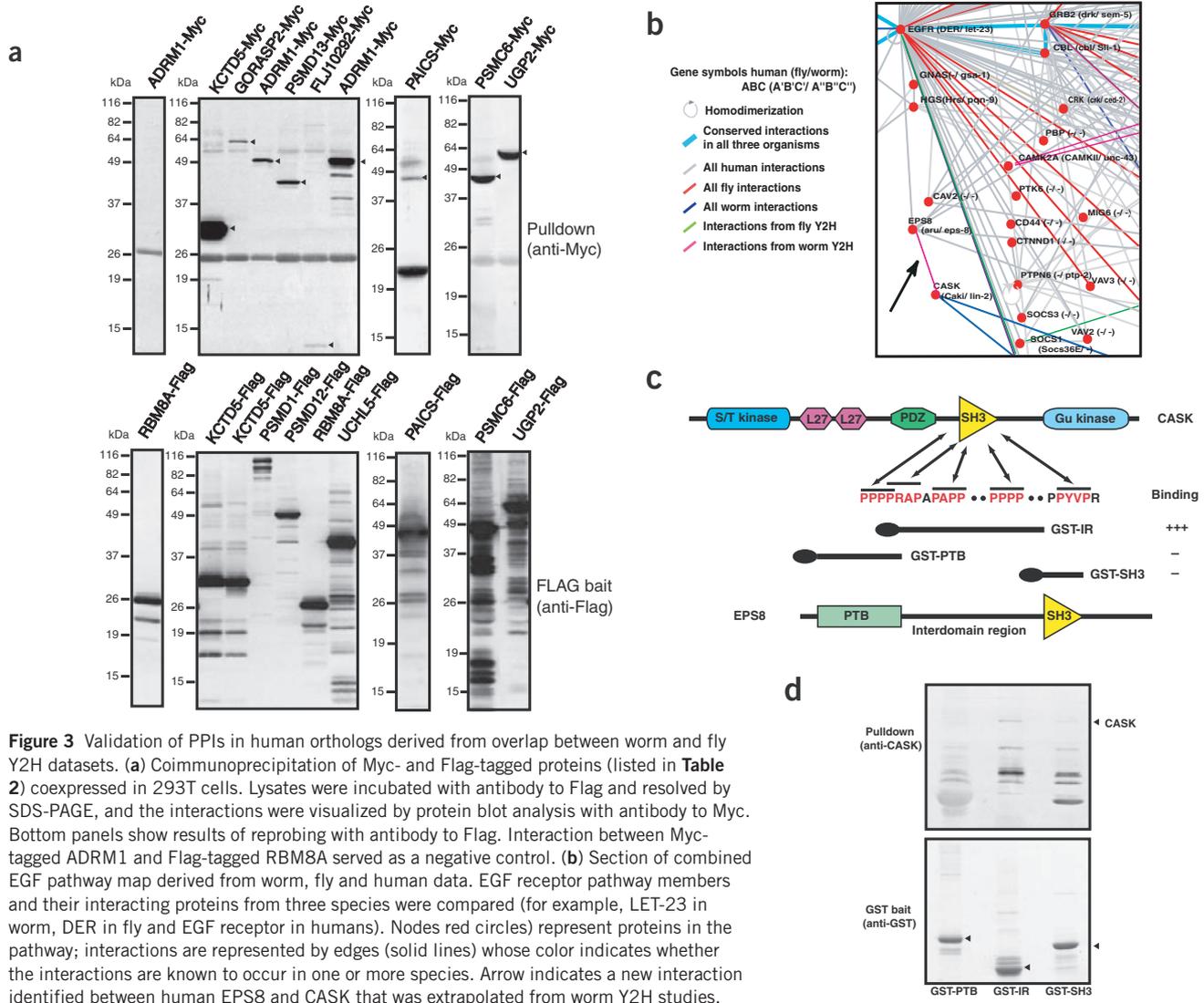
### Experimental validation of predicted human PPIs

To predict new PPIs, we used the overlap between the worm and fly Y2H datasets as an indicator of interactions that were most likely to occur in humans. Interspecies transfer of PPIs from *Helicobacter pylori* to *Escherichia coli* and *Campylobacter jejuni*<sup>23</sup> and from yeast, worm and fly to humans<sup>24</sup> has been reported previously. However, because of the high incidence of false positives in the Y2H system, our goal was to obtain a narrower set of PPIs from two independent screens in two different species—worm and fly—as there is a greater chance that those PPIs represent *bona fide* interactions.

Thirty-six interactions were detected in fly and worm (Fig. 2a) that have not yet been described in humans (Supplementary Table 2 online).

**Table 2** Experimentally validated PPIs in humans

Protein (function)	Protein (function)	Features of interaction
PAICS (phosphoribosylaminoimidazole carboxylase)	PAICS (phosphoribosylaminoimidazole carboxylase)	Homomeric
UGP2 (UDP-glucose pyrophosphorylase-2)	UGP2 (UDP-glucose pyrophosphorylase-2)	Homomeric
PSMC6 (proteasome 26S subunit, ATPase subunit-6)	PSMC6 (proteasome 26S subunit, ATPase subunit-6)	Homomeric; PSMC6 may link 19S regulatory particle with 'modulator' complex
KCTD5 (potassium channel tetramerization domain-containing-5)	KCTD5 (potassium channel tetramerization domain-containing-5)	Homomeric
GORASP2 (Golgi reassembly stacking protein-2)	KCTD5 (potassium channel tetramerization domain-containing-5)	Heteromeric; interaction may assist in transport and targeting of KCTD5.
ADRM1 (adhesion-regulating molecule-1)	UCHL5 (ubiquitin carboxyl-terminal hydrolase L5)	Heteromeric; interaction may regulate proteolysis of ADRM1 by 26S proteasome.
PSMD13 (proteasome 26S, non-ATPase subunit-13)	PSMD12 (proteasome 26S, non-ATPase subunit-12)	Heteromeric; first interaction identified among non-ATPase subunit in human 19S proteasome-regulatory particles; also observed in <i>S. cerevisiae</i>
FLJ10292 (unknown function)	RBM8A (RNA-binding motif protein-8A)	Heteromeric
ADRM1 (adhesion-regulating molecule-1)	PSMD1 (proteasome 26S, non-ATPase subunit-1)	Heteromeric



**Figure 3** Validation of PPIs in human orthologs derived from overlap between worm and fly Y2H datasets. **(a)** Coimmunoprecipitation of Myc- and Flag-tagged proteins (listed in **Table 2**) coexpressed in 293T cells. Lysates were incubated with antibody to Flag and resolved by SDS-PAGE, and the interactions were visualized by protein blot analysis with antibody to Myc. Bottom panels show results of reprobing with antibody to Flag. Interaction between Myc-tagged ADRM1 and Flag-tagged RBM8A served as a negative control. **(b)** Section of combined EGF pathway map derived from worm, fly and human data. EGF receptor pathway members and their interacting proteins from three species were compared (for example, LET-23 in worm, DER in fly and EGF receptor in humans). Nodes (red circles) represent proteins in the pathway; interactions are represented by edges (solid lines) whose color indicates whether the interactions are known to occur in one or more species. Arrow indicates a new interaction identified between human EPS8 and CASK that was extrapolated from worm Y2H studies. **(c)** Schematic depiction of domain structures of EPS8, CASK and GST-EPS8 fusion protein. The five PXXP motifs in the interdomain region of EPS8 are marked and are likely to mediate its interaction with the SH3 domain of CASK. **(d)** Pulldown experiment to detect association of CASK and EPS8. HeLa cell lysates were incubated with GST-PTB, GST-IR or GST-SH3 fusion proteins derived from EPS8 protein. Bound CASK was visualized by protein blotting with antibody to CASK (top), and GST fusion proteins on nitrocellulose membrane were visualized by MemCode protein stain (bottom).

Only 21 of these 36 interactions were described as being high confidence in both the worm and fly datasets. We carried out coimmunoprecipitation experiments to validate 9 of these 36 interactions in humans. Because antibodies against most of these protein pairs were not commercially available, we used epitope-tagged cDNA constructs in human embryonic kidney 293T cells. Full-length cDNAs for human orthologs of the protein pairs were obtained and tagged with Flag or Myc epitopes. We were able to validate all nine interactions in this fashion (see **Table 2** for a list of proteins). **Figure 3a** shows protein blots to detect the interacting proteins in anti-Flag coimmunoprecipitates using antibody to Myc.

The interactions verified by these experiments included a homomeric interaction between four gene products: PAICS, an enzyme whose expression level is modulated in certain cancers<sup>25,26</sup>; UGP2, an enzyme that catalyzes the transfer of a glucose moiety from glucose 1-phosphate to MgUTP; KCTD5, a new member of a family of voltage-gated potassium channel complexes; and, PSMC6, an ATPase subunit. Even though

none of these proteins dimerize or form higher-order structures, their functions are likely to be regulated in this fashion. As a negative control, we found that ADRM1 does not interact with RBM8A (**Fig. 3a**).

We detected heteromeric interactions among five protein pairs: KCTD5 and GORASP2, FLJ10292 and RBM8A, PSMD13 and PSMD12, ADRM1 and UCHL5, and ADRM1 and PSMD1. GORASP2 is a component of the Golgi stacking machinery<sup>27</sup> that has an important role in maintaining Golgi structure and protein transport<sup>28</sup>. FLJ10292 is closely related to MAGOH, the human homolog of the fly protein Magonashi, and human MAGOH and RBM8A are components of the exon-exon junction complex<sup>29</sup>. The 26S proteasome is composed of three complexes, one 20S catalytic particle and two identical 19S regulatory particles. Although the interactions among the human 19S base components have been extensively studied<sup>30-32</sup>, those among non-ATPase subunits of the lid have not been well studied. The interaction between PSMD13 and PSMD12 shown in **Figure 3a** is the first validated interaction among the non-ATPase subunits in the lid of the human 19S

regulatory particle. A 'modulator' complex that enhances the catalytic activity of the 19S subunit was recently isolated<sup>33</sup>. It is not known how this complex, which consists of PSMC3, PSMD9 and PSMC6, associates with the 19S regulatory particles. However, our finding that PSMC6 forms homo-oligomers strongly suggests that the association between the modulator complex and the 19S regulatory particle is mediated by PSMC6, as it is a component of both of these complexes. This hypothesis is supported by the presence of a coiled-coil region in PSMC6 that could lead to a homomeric interaction. The last two interactions that we validated involved ADRM1, a type I transmembrane protein that possesses three dileucine motifs and is involved in trafficking and degradation of a variety of transmembrane proteins, including the insulin receptor<sup>34</sup>. The interaction of ADRM1 with UCHL5, a deubiquitinating enzyme, may contribute to downregulation of cell adhesion by targeting ADRM1 for degradation. ADRM1 also binds to PSMD1, a component of the 19S proteasomal subunit.

An examination of 7 of the 27 remaining interactions that overlapped revealed that they were previously reported as components of protein complexes in humans (as opposed to direct interactions), and it is likely that they bind directly to each other in these complexes. It is notable that several of these PPIs might not have been pursued further because they were not reported as being high confidence. Thus, it is crucial to compare the data from a given Y2H experiment with interaction data on orthologous proteins from different species to identify those interactions that might be real but do not attain a high-confidence value in one particular dataset.

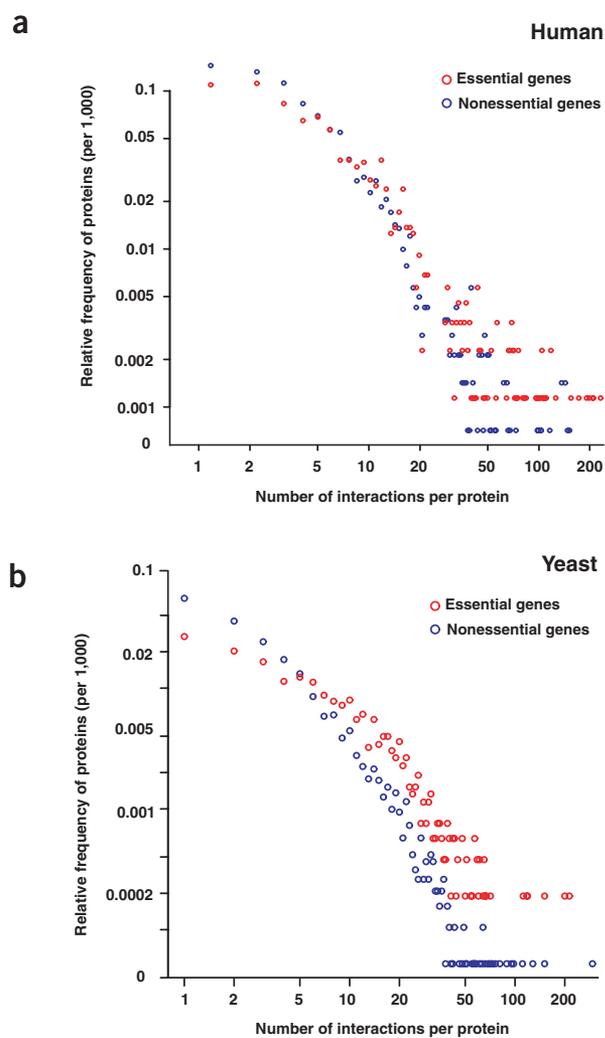
### Cross-species pathway analysis

We undertook a pathway-based approach to identify new interactions in the epidermal growth factor (EGF) pathway, which is conserved from nematodes to humans. *C. elegans* and *D. melanogaster* have been useful models for studying EGF signaling. For this analysis, the fly and worm orthologs of human proteins involved in the EGF receptor pathway were taken from the literature and three different databases (The Interactive Fly, FlyBase and WormBase). Our cross-species analysis revealed a potential interaction between EPS8 and CASK, two components of the EGF receptor pathway (Fig. 3b). This interaction was identified by Y2H assay in the worm<sup>8</sup> but has not been described in the fly or human. EPS8 is a tyrosine-phosphorylated substrate of EGF receptor that is involved in the transfer of signals between Ras and Rac by forming a complex with E3B1 and SOS1, thereby regulating actin remodeling<sup>35</sup>. The orthologs of CASK in the fly (*caki*) and worm (*lin-2*) are involved in the localization of EGF receptor through a complex with homologs of VELI1 and APBA1. We tested the prediction that human EPS8 and CASK would associate with each other.

Figure 3c shows the domain structure of the two proteins. The interdomain region of EPS8, but not the PTB or SH3 domains, specifically associated with CASK (Fig. 3d). Thus, this interaction is probably mediated by one or more of the five PXXP sequences (where 'x' represents any amino acid) in the interdomain region of EPS8 (Fig. 3c). This interaction of CASK with EPS8 might be involved in cytoskeletal organization mediated by EGF receptor signaling. This example shows the power of a comparative pathway analysis approach to identify potential interactions from data already available in other species.

### Essentiality of genes and connectivity

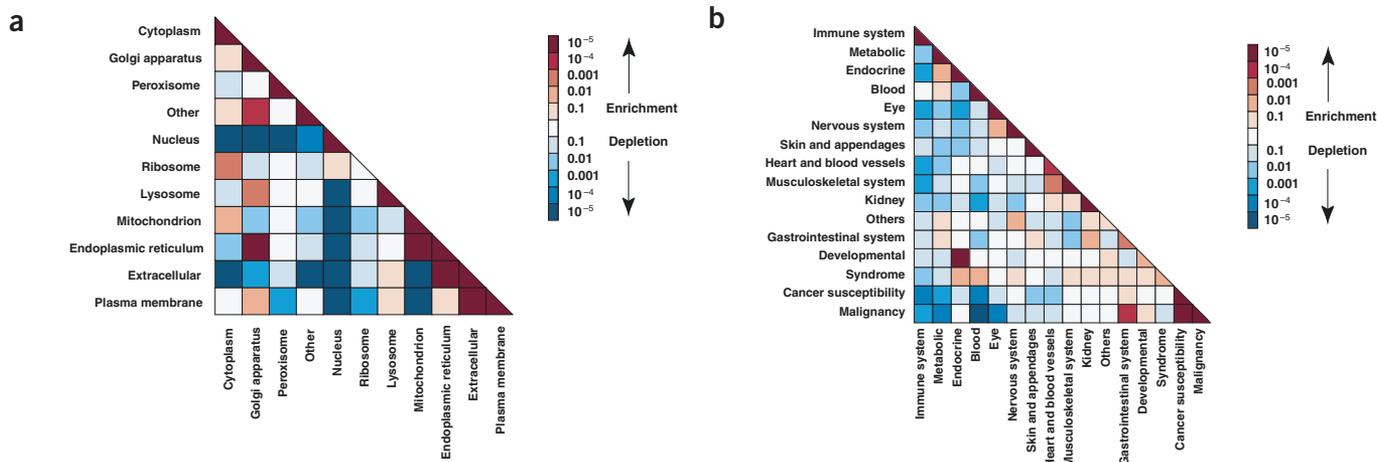
On the basis of a comparison of PPIs among essential and nonessential genes in yeast, it has been suggested that essential genes are the most highly interconnected proteins and resemble 'hubs'<sup>36</sup>. This prompted us to examine whether the same holds true in mammals, which, because of their complexity and multicellularity, might be expected to have a



**Figure 4** Distribution of PPIs for essential or nonessential genes (log<sub>10</sub> scale). (a) Human dataset. (b) Yeast dataset.

larger cohort of essential genes than yeast. On the other hand, if there is functional redundancy, one might expect a smaller cohort of essential genes. To address these issues, we cataloged genes whose ablation in mice resulted in either a lethal or a viable phenotype. We obtained information for 2,284 mouse genes: 1,359 from the curated knockout collection in the Mouse Genome Database<sup>37</sup> and an additional 925 from our own literature search using PubMed. Of these knockouts, 877 genes yielded lethal phenotypes, indicating that they are absolutely essential for survival.

Because a comprehensive dataset for yeast knockouts is now available, we also re-examined the connection between connectivity and lethality of genes in yeast. The previous yeast analysis was carried out on 1,870 genes in 2001, by which time only 2,240 interactions had been reported. In contrast, we now have lethality data on 6,014 yeast genes and 16,069 interactions, permitting a more comprehensive evaluation. A plot of the relative frequency of occurrence of proteins as a function of the number of interactions per protein for essential and nonessential genes showed that although statistically significant ( $P < 0.001$  by Wilcoxon<sup>38</sup> unpaired test), the difference in distributions in both humans (Fig. 4a) and yeast (Fig. 4b) is not pronounced. We also observed that the lethality of a gene could not be confidently predicted on the basis of the number of interaction



**Figure 5** Correlation of PPIs with subcellular localization and disease association. **(a)** Correlation of subcellular localization of proteins with PPIs. Shown is a matrix of the distribution of PPIs, where each axis represents the main site of subcellular localization (according to the HPRD) of one of the proteins in an interaction pair. As the matrix is symmetrical, only the lower diagonal is shown. Colors indicate fold enrichment or depletion calculated as the ratio of the number of edges in the observed network to the expected number of edges in an ensemble of random networks (see Methods, enrichment analysis). **(b)** Correlation of disease categories with PPIs. Each axis represents the category of disease associated with the gene encoding one of the proteins in an interaction pair.

partners or by an analysis of the receiver-operating characteristics curve (Supplementary Fig. 4 online). As has been recently discovered in the case of yeast<sup>39</sup>, we found that the current PPI data in humans and yeast do not provide evidence of a clear relationship between lethality and connectivity of genes. However, it is conceivable that a greater number of interactions could provide a clearer picture.

### Subcellular localization and PPIs

Interacting proteins are expected to have the same subcellular localization. We therefore examined the human interactome for enrichment or depletion of interactions in which both partners were localized to the same subcellular compartment. There was a statistically significant enrichment of interactions for almost all subcellular locations (Fig. 5a), indicating that interactions generally occurred between proteins residing in the same subcellular compartment. The number of interactions as a function of the subcellular localization of interacting proteins is shown in Table 3. The notable exceptions were cases where extracellular proteins interacted with those on the plasma membrane or in the endoplasmic reticulum and where proteins known to reside in the Golgi apparatus interacted with those in the endoplasmic reticulum. In contrast, we did not see interaction of nuclear or mitochondrial proteins with those in most other compartments. A closer inspection of the other interactions that were enriched across subcellular compartments (endoplasmic reticulum–mitochondrion, endoplasmic reticulum–extracellular) showed that in numerous cases, at least one of the interacting proteins was alternatively localized to a compartment other than its major site of localization (and thus the interacting proteins did indeed share a common compartment). This is not an uncommon situation, as >28% of proteins annotated with subcellular localization information in the HPRD are, in fact, localized to two or more compartments (>8% are localized to three or more compartments).

### Interactions of gene products involved in disease

Because proteins are generally part of larger protein complexes, it is reasonable to suppose that human diseases which share a common biochemical mechanism could be caused by mutations in one of sev-

eral genes that interact at the protein level in a pathway or subnetwork. In support of this, we found that the genes ascertained from the Online Mendelian Inheritance in Man (OMIM) database<sup>40</sup> to be associated with a human disease preferentially interacted with other disease-causing genes ( $P = 6 \times 10^{-23}$ ) over those without any known disease association.

We next assessed whether genes that cause the same type of diseases are more likely to interact with each other. For this, we classified the OMIM disease entries into categories based on the organ system affected (the nervous system, for example) or the type of disorder (metabolic disorders, for example) and carried out an enrichment analysis on the 2,850 binary interactions in which both interacting proteins were derived from disease-associated genes (a total of 1,077 genes linked to 3,133 disease entities). Although this classification is somewhat arbitrary, we observed a strong enrichment for interactions within the majority of disease categories, including immune system, metabolic, blood, eye and nervous system disorders and malignancies (Fig. 5b). One of the strongest enrichments across categories was between the endocrine and developmental disorders. Because of this interconnectivity among disease-causing genes and, more specifically, within disease categories, it should be possible to identify additional candidate genes for diseases through this network approach, as suggested recently<sup>41</sup>. Although additional analyses will be required, these studies suggest that the proteins encoded by genes that mutate to cause a detectable phenotype or a disease are involved in a pathway or in signal transduction mechanisms. Such a network- and pathway-oriented approach, coupled with information on the chromosomal location of suspected loci, transcriptomic profiles and tissue predilection of diseases, might help refine candidate gene lists in future association and linkage studies.

### DISCUSSION

Our analysis of the human interactome has confirmed some previous expectations and refuted or cast doubt on others. The lack of overlap between datasets from different species is notable and cannot be attributed to the use of only high-confidence data from Y2H studies or to the lack of orthologous pairs of proteins in datasets. In spite of the low

overlap, the overlaps that were identified were strong predictors of true interactions in other species. We did not find any strong evidence for an association between the connectivity of genes and their lethality, despite the fact that our datasets were much larger than those used in previous analyses<sup>36</sup>.

Complex networks show a high degree of clustering. The clustering coefficient,  $C_i$ —the ratio of the number of links between a node's neighbors to the number of interactions in the node—provides a neighborhood measure of proteins and reflects the structural integrity of a network<sup>42</sup>. The human interaction network has an average  $C_i$  of 0.10, which compares well with those of the yeast (0.11) and worm (0.08) networks. The  $C_i$  for the fly interaction dataset is much lower (0.02). The higher connectivity of the yeast and human networks probably reflects the fact that PPIs in these two organisms have been systematically elucidated by biochemical experiments over several decades (in addition to the high-throughput experiments in yeast). The lower connectivity of the fly dataset<sup>5</sup> could be explained by the fact that most of the interactions were obtained by screening against the entire library instead of a subset of ORFs (which was done in the case of the worm Y2H study<sup>8</sup>).

As expected, we found that most of the interacting protein pairs colocalize to the same subcellular compartment. Our data also revealed a small number of interactions between proteins localized to different subcellular locations. These might represent experimental errors in the original data. However, evidence derived from literature-based annotations indicates that it is more likely to result from proteins having more than one subcellular location. We also showed that proteins encoded

by disease-associated genes are highly connected and that genes causing similar diseases form subnetworks. This interconnectivity could be used in genome-wide searches for identification of candidate disease genes.

Although the confidence level of binary interactions in individual high-throughput analyses is low, it is possible to predict interactions in other species more confidently if the interactions that overlap in multiple species are considered. Alternatively, if there are highly connected subclusters observed in a single species, it might also be possible to predict new components in similar subclusters in other species. The human interaction data presented here should provide a platform for testing predictions from other types of data and generating new hypotheses.

## METHODS

**Interaction datasets and protein annotations.** We obtained 24,385 literature-derived human PPIs from the HPRD<sup>10</sup>. Other human PPI data included interactions from BIND<sup>11</sup> (4,644 interactions), DIP<sup>12</sup> (1,049 interactions), MIPS<sup>13</sup> (334 interactions), MINT<sup>14</sup> (3,544 interactions) and IntAct<sup>15</sup> (2,420 interactions). The combined data from datasets other than HPRD were first verified by checking with the literature. This procedure resulted in a total of 25,464 unique PPIs that could be mapped to unique gene identifiers.

The yeast interaction dataset was downloaded from DIP<sup>12</sup> and MIPS<sup>18</sup> (29 June 2005), the fly dataset was compiled by merging the data from refs. 5, 7 and 6, and the worm dataset was from ref. 8. The protein localization annotations were obtained from the HPRD. Disease annotations were from OMIM<sup>40</sup>. Protein interaction networks were generated using the Osprey network visualization system<sup>43</sup>.

**Table 3** Numbers of PPIs, by subcellular localization of interacting proteins

Cytoplasm (1,504)	3,002											
Golgi apparatus (133)	182	43										
Peroxisome (27)	16	—	10									
Other (88)	124	14	—	24								
Nucleus (1,756)	2,556	45	3	67	4,140							
Ribosome (58)	98	1	—	—	74	—						
Lysosome (33)	29	6	—	1	10	1	6					
Mitochondrion (188)	270	3	—	1	102	—	—	114				
ER (184)	191	23	—	2	69	2	2	30	66			
Extracellular (611)	247	15	2	3	132	10	10	17	91	670		
Plasma membrane (1,367)	2,383	113	2	63	484	22	22	77	137	927	1,528	
	Cytoplasm (1,504)	Golgi apparatus (133)	Peroxisome (27)	Other (88)	Nucleus (1,756)	Ribosome (58)	Lysosome (33)	Mitochondrion (188)	ER (184)	Extracellular (611)	Plasma membrane (1,367)	

The numbers in parentheses indicate the total number of proteins in the interactome that are present in each subcellular compartment (based on data from HPRD<sup>10</sup>). ER, endoplasmic reticulum.

**Assignment of orthologs.** The ortholog mapping dataset was downloaded from the InParanoid database<sup>19</sup> (version 4.0), which considered 22,216 genes in humans (73% of the human interactome), 5,777 genes in yeast (98% of the yeast interactome), 19,765 in worm (98% of the worm interactome) and 13,792 in fly (97% of the fly interactome). Only the orthologous sequence pairs with a bootstrap value of 100% and a score of 1.00 in each cluster were considered for this analysis.

**Lethality datasets.** We obtained information for a total of 2,284 mouse genes. Of these, 1,359 were obtained from the curated knockout collection of the Mouse Genome Database<sup>37</sup> and an additional 925 were obtained from our literature search of PubMed. Embryonic, neonatal, perinatal and postnatal lethality were combined into a single category of genes causing lethal phenotypes. The yeast knockout data were obtained from the Saccharomyces Genome Database<sup>44</sup>.

$$C_i = \frac{\sum_{i=1 \rightarrow N} \frac{2e_i}{k_i(k_i-1)}}{N}$$

**Calculation of  $C_i$ .** The degree of connectivity of human, worm and fly interaction networks was determined by calculating the  $C_i$  using the formula where  $k_i$  is the degree of node  $i$ ,  $i$  is the  $i^{\text{th}}$  node,  $N$  is the total number of nodes, and  $e_i$  is the number of interactions existing between the  $k_i$  protein connected to protein  $i$ .

**Coimmunoprecipitation and GST pulldown experiments.** For coimmunoprecipitation studies, cDNAs were amplified and cloned into mammalian expression vectors containing sequences for Myc or Flag epitopes. The construct pairs to be tested for interaction were cotransfected into 293T cells using the calcium phosphate method as previously described<sup>45</sup>. GST-deletion constructs of EPS8 were made, and the proteins were purified and used for binding studies as described previously<sup>45</sup>.

**Statistical analyses.** We tested the null hypothesis that there is no difference in the number of interactions for essential and nonessential genes against the alternative hypothesis that any difference may exist. We used the Wilcoxon unpaired test<sup>38</sup>, a nonparametric test that is appropriate for this application because of the highly skewed distributions and subtle patterns of differences among distributions.

To assess whether it is possible to predict the essentiality of a gene on the basis of the number of interactions alone, we generated receiver-operating characteristics curves and positive predictive values<sup>46</sup>. We computed positive predictive values by restricting the analysis to proteins that have at least  $x$  interactions for a range of values of  $x$ . Confidence intervals for these values were computed by a normal approximation of the binomial proportion<sup>47</sup>. Statistical analyses were performed using the R language and software<sup>48</sup>.

**Enrichment analysis.** We used a randomization algorithm to characterize and assess the statistical significance of the enrichment of types of edges in the network of protein interactions. For notation, we used indices to represent the

$$i = 1, 2, K, N$$

proteins in the network and  $e_{ij} = e_{ji} = 1/0$  to indicate whether or not a pair of proteins is connected by an edge. The indices represent annotation of a protein

$$\alpha = 1, 2, \dots, K$$

in one of  $K$  total, non-mutually exclusive categories, with  $c_{i\alpha} = 1/0$  depending on whether protein  $i$  is or is not annotated as a member of category  $\alpha$ . To avoid possibly spurious enrichment, we deleted self-interactions before testing for enrichment.

For each pair of categories, the number of edges observed in the network is

$$n_{\alpha\beta}^{\text{obs}} = \sum_{i < j} (c_{i\alpha} c_{j\beta} \text{ OR } c_{i\beta} c_{j\alpha}) e_{ij}$$

Thus, if connected proteins are each annotated as belonging to both categories  $\alpha$  and  $\beta$ , this counted as only a single  $\alpha\beta$  edge. Next, the probability distribution

of  $n_{\alpha\beta}$  was calculated for the ensemble of random networks that maintained the protein annotations, the expected degree  $k_i$  of each protein and the total expected number of edges.

$$E = (1/2) \sum_i k_i$$

The mean of  $n_{\alpha\beta}$  is

$$\bar{n}_{\alpha\beta} = \sum_{i < j} (c_{i\alpha} c_{j\beta} \text{ OR } c_{i\beta} c_{j\alpha}) \bar{e}_{ij}$$

where  $\bar{e}_{ij}$  is the expectation in a random network of an edge between proteins  $i$  and  $j$ ,  $(k_i k_j / 2E) / (1 + k_i k_j / 2E)$ . Similarly, the variance of  $n_{\alpha\beta}$  is

$$\text{var}(n_{\alpha\beta}) = \sum_{i < j} (c_{i\alpha} c_{j\beta} \text{ OR } c_{i\beta} c_{j\alpha}) \bar{e}_{ij} (1 - \bar{e}_{ij})$$

Fold enrichment was calculated as

$$n_{\alpha\beta}^{\text{obs}} / \bar{n}_{\alpha\beta}$$

We noted that

$$\bar{n}_{\alpha\beta} \approx \text{var}(n_{\alpha\beta})$$

for all annotations, as is expected for a sparse network, which permits use of a Poisson distribution to calculate statistical significance:

$$P \text{ value}(n) = \begin{cases} \sum_{j=0}^n \bar{n}^j \exp(-\bar{n}) / j! , & n < \bar{n} \\ \sum_{j=n}^{\infty} \bar{n}^j \exp(-\bar{n}) / j! , & n \geq \bar{n} \end{cases}$$

The index ( $\alpha\beta$ ) is suppressed for simplicity. Finally, we applied a multiple-testing correction as  $P \text{ value}(\text{multi}) = 1 - (1 - P)^m$ , where  $P$  is the single-test  $P$  value and  $m$  is the number of tests, which was tabulated separately for tests of enrichment ( $m$  = number of  $\alpha\beta$  pairs with at least one edge in the observed network) and depletion ( $m$  = number of  $\alpha\beta$  pairs possible in the randomized networks).

**URLs.** The human PPI data used for analysis in this manuscript are available in PSI-MI format for download at [http://www.hprd.org/2006/Nature\\_Genetics](http://www.hprd.org/2006/Nature_Genetics). The most recent version of HPRD data is available at <http://www.hprd.org/download>.

*Note: Supplementary information is available on the Nature Genetics website.*

#### ACKNOWLEDGMENTS

A.P., J.D.B. and J.S.B. were supported by a grant from the National Institutes of Health (U54 RR020839). G.P. was supported by a grant from the National Science Foundation (NSF 034211), J.S.B. was supported by grants from the US National Institutes of Health (R41 GM073492 and R01 GM067761) and the Whitaker Foundation and S.P. was supported by the IZKF Würzburg project. The authors thank J. Eppig and C. Bult for providing knockout data from the Mouse Genome Database and B. Migeon, N. Katsanis and J. Mendell for helpful suggestions. The HPRD was developed with funding from the National Institutes of Health and the Institute of Bioinformatics.

#### COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Published online at <http://www.nature.com/naturegenetics>  
Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

- Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
- Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein

- interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
3. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
  4. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
  5. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
  6. Formstecher, E. *et al.* Protein interaction mapping: a *Drosophila* case study. *Genome Res.* **15**, 376–384 (2005).
  7. Stanyon, C.A. *et al.* A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol.* **5**, R96 (2004).
  8. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
  9. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
  10. Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
  11. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
  12. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, Database issue, D449–D451 (2004).
  13. Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832–834 (2005).
  14. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
  15. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
  16. Matthews, L.R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “interologs”. *Genome Res.* **11**, 2120–2126 (2001).
  17. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* **102**, 1974–1979 (2005).
  18. Guldener, U. *et al.* CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.* **33**, D364–D368 (2005).
  19. O’Brien, K.P., Remm, M. & Sonnhammer, E.L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).
  20. Hazbun, T.R. & Fields, S. Networking proteins in yeast. *Proc. Natl. Acad. Sci. USA* **98**, 4277–4278 (2001).
  21. Legrain, P. & Selig, L. Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett.* **480**, 32–36 (2000).
  22. Mrowka, R., Patzak, A. & Herzel, H. Is there a bias in proteome research? *Genome Res.* **11**, 1971–1973 (2001).
  23. Wojcik, J., Boneca, I.G. & Legrain, P. Prediction, assessment and validation of protein interaction maps in bacteria. *J. Mol. Biol.* **323**, 763–770 (2002).
  24. Lehner, B. & Fraser, A.G. A first-draft human protein-interaction map. *Genome Biol.* **5**, R63 (2004).
  25. Sun, W. *et al.* Identification of differentially expressed genes in human lung squamous cell carcinoma using suppression subtractive hybridization. *Cancer Lett.* **212**, 83–93 (2004).
  26. Zaza, G. *et al.* Acute lymphoblastic leukemia with TEL-AML1 fusion has lower expression of genes involved in purine metabolism and lower de novo purine synthesis. *Blood* **104**, 1435–1441 (2004).
  27. Shorter, J. *et al.* GRASP55, a second mammalian GRASP protein involved in the stacking of Golgi cisternae in a cell-free system. *EMBO J.* **18**, 4949–4960 (1999).
  28. Short, B. *et al.* A GRASP55-rab2 effector complex linking Golgi structure to membrane traffic. *J. Cell Biol.* **155**, 877–883 (2001).
  29. Kataoka, N., Diem, M.D., Kim, V.N., Yong, J. & Dreyfuss, G. Magoh, a human homolog of *Drosophila mago nashi* protein, is a component of the splicing-dependent exon-exon junction complex. *EMBO J.* **20**, 6424–6433 (2001).
  30. Gorbea, C., Taillandier, D. & Rechsteiner, M. Mapping subunit contacts in the regulatory complex of the 26 S proteasome. S2 and S5b form a tetramer with ATPase subunits S4 and S7. *J. Biol. Chem.* **275**, 875–882 (2000).
  31. Ishizuka, T. *et al.* Human immunodeficiency virus type 1 Tat binding protein-1 is a transcriptional coactivator specific for TR. *Mol. Endocrinol.* **15**, 1329–1343 (2001).
  32. Richmond, C., Gorbea, C. & Rechsteiner, M. Specific interactions between ATPase subunits of the 26 S protease. *J. Biol. Chem.* **272**, 13403–13411 (1997).
  33. Fujiwara, T., Watanabe, T.K., Tanaka, K., Slaughter, C.A. & DeMartino, G.N. cDNA cloning of p42, a shared subunit of two proteasome regulatory proteins, reveals a novel member of the AAA protein family. *FEBS Lett.* **387**, 184–188 (1996).
  34. Haft, C.R., Klausner, R.D. & Taylor, S.I. Involvement of dileucine motifs in the internalization and degradation of the insulin receptor. *J. Biol. Chem.* **269**, 26286–26294 (1994).
  35. Scita, G. *et al.* EPS8 and E3B1 transduce signals from Ras to Rac. *Nature* **401**, 290–293 (1999).
  36. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
  37. Eppig, J.T. *et al.* The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res.* **33**, 471–475 (2005).
  38. Hollander, M. & Wolfe, D.A. *Nonparametric Statistical Inference* 27–33 (John Wiley & Sons, New York, 1973).
  39. Coulomb, S., Bauer, M., Bernard, D. & Marsolier-Kergoat, M.C. Gene essentiality and the topology of protein interaction networks. *Proc. Biol. Sci.* **272**, 1721–1725 (2005).
  40. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, Database issue, D514–D517 (2005).
  41. Arking, D.E., Chugh, S.S., Chakravarti, A. & Spooner, P.M. Genomics in sudden cardiac death. *Circ. Res.* **94**, 712–723 (2004).
  42. Albert, R. & Barabasi, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
  43. Breitkreutz, B.J., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biol.* **4**, R22 (2003).
  44. Cherry, J.M. *et al.* SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* **26**, 73–79 (1998).
  45. Pandey, A. *et al.* Cloning and characterization of PAK5, a novel member of mammalian p21-activated kinase-II subfamily that is predominantly expressed in brain. *Oncogene* **21**, 3939–3948 (2002).
  46. Egan, J.P. *Signal Detection Theory and ROC Analysis* (Academic, New York, 1975).
  47. Cox, D.R. & Snell, E.J. *Analysis of Binary Data* (Chapman and Hall, London, 1970).
  48. Ihaka, R. & Gentleman, R.R. A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5**, 299–314 (1996).