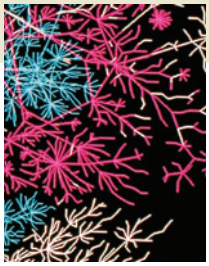

Probabilistic model of the human protein-protein interaction network

Daniel R Rhodes[1,2,7], Scott A Tomlins[2,7], Sooryanarayana Varambally[2,7], Vasudeva Mahavisno[2], Terrence Barrette[2], Shanker Kalyana-Sundaram[2], Debashis Ghosh[3], Akhilesh Pandey[6] & Arul M Chinnaiyan[1,2,4,5]

A catalog of all human protein-protein interactions would provide scientists with a framework to study protein deregulation in complex diseases such as cancer. Here we demonstrate that a probabilistic analysis integrating model organism interactome data, protein domain data, genome-wide gene expression data and functional annotation data predicts nearly 40,000 protein-protein interactions in humans—a result comparable to those obtained with experimental and computational approaches in model organisms. We validated the accuracy of the predictive model on an independent test set of known interactions and also experimentally confirmed two predicted interactions relevant to human cancer, implicating uncharacterized proteins into definitive pathways. We also applied the human interactome network to cancer genomics data and identified several interaction subnetworks activated in cancer. This integrative analysis provides a comprehensive framework for exploring the human protein interaction network.

We began by assembling a collection of genomic and proteomic data potentially useful in predicting human protein-protein interactions that included model organism protein-protein interactions[1], protein domain assignments[2], gene expression measurements in human tissue samples[3] and biological function annotations[4] (**Table 1**). Based on previous reports, we suspected that (i) model organism interactions may suggest interactions among orthologous human proteins[5,6], (ii) similar gene expression profiles across a panel of human tissue samples may identify interacting protein products[7,8], (iii) protein domain pairs enriched among known human protein-protein interactions may suggest novel interactions[9], (iv) shared functional annotations from Gene Ontology[4] may suggest physical interactions, and (v) that combining evidence from independent data sources may strongly predict protein-protein interactions[10–12]. To test these hypotheses, we applied a naive Bayes classifier[7], a method well-suited for integrating disparate data types.

[1]Bioinformatics Program, Departments of [2]Pathology, [3]Biostatistics, [4]Urology and [5]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [6]Mc-Kusick-Nathans Institute of Genetic Medicine and the Department of Biological Chemistry, Johns Hopkins University, Baltimore, Maryland 21205, USA. [7]These authors contributed equally to this work.
Correspondence should be addressed to A.M.C. (arul@umich.edu).

Published online 4 August 2005; doi:10.1038/nbt1103

A gold standard positive set (GSP) of 11,678 distinct protein-protein interactions among 5,505 proteins was queried from the Human Protein Reference Database (HPRD)[12], a resource that contains known protein-protein interactions manually curated from the literature by expert biologists. A gold standard negative set (GSN) of 3,106,928 protein pairs was defined, in which one protein was assigned to the plasma membrane cellular component and the other to the nuclear cellular component by the Gene Ontology Consortium[4]. Although it is known that membrane proteins can occasionally interact with nuclear proteins, we demonstrated that there are far fewer known interactions within GSN than would be expected by chance (**Supplementary Methods** online). By averaging the number of interactions per protein in the GSP, we estimated the prior odds of interaction among two randomly selected proteins to be 1 in 381. This is likely an underestimate of the true prior odds because all protein-protein interactions are not known; however, to err on the conservative side, we assumed the estimate to be true. To achieve posterior odds (O_{post}) greater than 1 (that is, a >50% chance of interaction), the likelihood ratio cutoff (LR_{cut}) must be set at 381. In the following sections, we systematically test the predictive data sets against the GSP and GSN, generating likelihood ratios, and then we combine the data set–specific likelihood ratios (LR) in a naive Bayes model, which is applied to all protein pairs to predict novel human protein-protein interactions. Contingency tables detailing the intersection of predicted interactions with the GSP and GSN sets and the resultant likelihood ratios are provided online as **Supplementary Tables 1–8**.

Model organism protein-protein interactions

From the Database of Interacting Proteins (DIP)[1], we queried high-throughput interactome data from three model organisms: *Sacchromyces cerevisiae*[13–16], *Caenorhabditis elegans*[17] and *Drosophila melanogaster*[3]. The *S. cerevisiae* interactome (SC) data comprised four high-throughput interactome data sets[13–16] and several low-throughput experiments, whereas the *D. melanogaster* (DM) and *C. elegans* (CE) data comprised one yeast two-hybrid data set each[17,18]. Human interactions were predicted by mapping model organism proteins to human orthologs using the Inparanoid database[19] (**Fig. 1a** and **Supplementary Table 1** online). The SC data had 10,200 interactions (13,134 entries) among 5,339 proteins, of which 2,580 mapped to 6,854 human proteins, predicting 20,405 orthologous human interactions. Of 775 possible GSPs, we predicted 256 (33.0%) to interact, in contrast to just 154 of 77,934 (0.20%) GSNs (LR = 167.2). The DM data had 20,709 interactions among 5,020 proteins, which mapped

Table 1 Data sources integrated to predict human protein-protein interactions

Prediction type	Data set	Number of data elements	Number of protein pairs
Ortholog interactions	*S. cerevisiae*	13,134	20,405
	C. elegans	4,008	2,895
	D. melanogaster	20,709	17,109
Coexpression	Breast	2,829,762	73,790,861
	Soft tissue	1,015,173	30,034,750
	Multi-cancer	2,148,204	41,550,914
	Liver	4,340,501	49,508,820
	Lymphoma	2,167,907	9,498,975
Shared biological process	GO	94,045	40,018,019
Enriched domain pair		19,438	531,489

Gold standard	Data set	Number of data elements	Number of protein pairs
Positive (GSP)	HPRD interactions	17,462	17,462
	Training set	11,678	11,678
	Test set	5,784	5,784
Negative (GSN)	GO, plasma membrane	1,397	3,106,928
	GO, nucleus	2,224	

to 6,864 human proteins, predicting 17,109 orthologous human interactions. Of 1,805 possible GSPs, we predicted 118 (6.5%) to interact, in contrast to 392 of 199,640 (0.19%) GSNs (LR = 33.3). Finally, the CE data had 4,008 interactions among 2,100 distinct proteins, which mapped to 1,181 human proteins, predicting 2,895 orthologous human interactions. Of 144 possible GSPs, we predicted 12 (8.3%) to interact in contrast to 49 of 12,222 (0.40%) GSNs (LR = 20.8). In summary, model organism interactions are moderately predictive of orthologous human protein interactions, although none of the model organism data sets reach the likelihood ratio threshold of 381. The SC data were found to be most predictive, likely owing to multiple semi-redundant experimental data sources as opposed to single data sources for DM and CE.

Next we examined parameters associated with the model organism interactions and tested their ability to stratify the predicted human interactions into confidence bins. These parameters included ortholog mapping confidence scores from Inparanoid, the number of human interactions predicted per model organism interaction, the number of evidence types in the case of the SC data, and an interaction confidence score in the case of the DM data. To derive logical bins associated with these parameters, we used a decision tree algorithm[20]. The interactions predicted from the SC data were grouped into four bins: a high-confidence bin, in which the SC interactions had more than one evidence type (for example, yeast two-hybrid and immunoprecipitation; $n = 3,248$; LR = 1,664.8), a medium-high confidence bin, in which a single human interaction was predicted

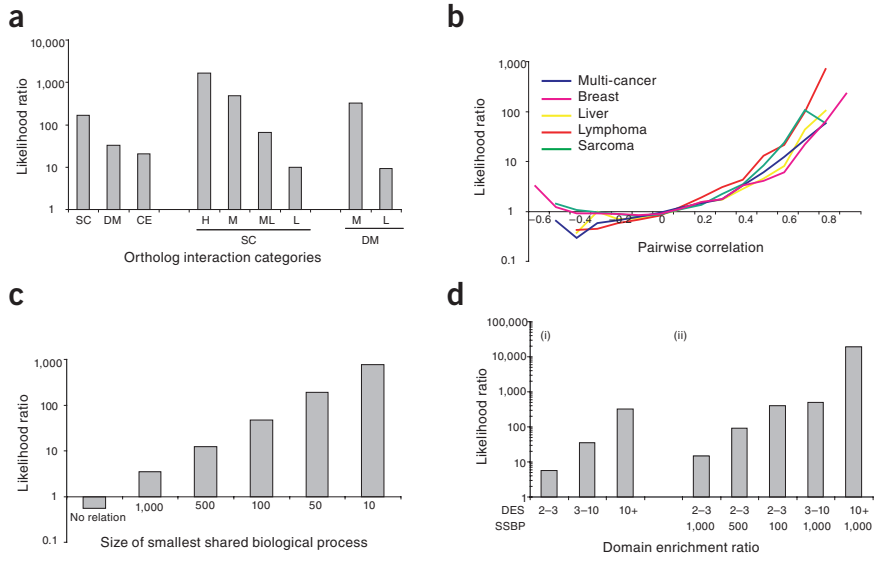

Figure 1 Diverse genomic and proteomic data sources contribute to the predictive modeling of human protein-protein interactions. (**a**) Model organism interaction data was downloaded from the DIP and model organism proteins were mapped to human orthologs using the Inparanoid database. All pairs of human proteins that corresponded to orthologs reported to interact were compared with the GSP and GSN interactions to generate likelihood ratios. A decision tree algorithm was used to bin predicted interactions into high (H), medium (M), and low (L) confidence groups when possible. (**b**) Coexpression predicts human protein-protein interactions. Five large gene expression data sets were selected from the Oncomine database based on predictive strength. Pairwise gene correlation matrices were calculated for each data set. (**c**) Shared biological function predicts human protein-protein interactions. For each pair of proteins, as a measure of functional similarity, the SSBP was identified and then gene pairs were binned based on this number. The likelihood of human protein-protein interactions increases as the size of the SSBP decreases. (**d**) Pairs of domains enriched among pairs of proteins known to interact were identified, ascribed a domain enrichment score (DES), binned by this score and tested against the GSP and GSN (i). Because the shared biological function and domain enrichment data sources were found to be somewhat redundant, protein pairs with both evidence sources were analyzed in conjunction; a decision tree algorithm grouped protein pairs with both evidence types into five bins based on the size of the SSBP and DES (ii).

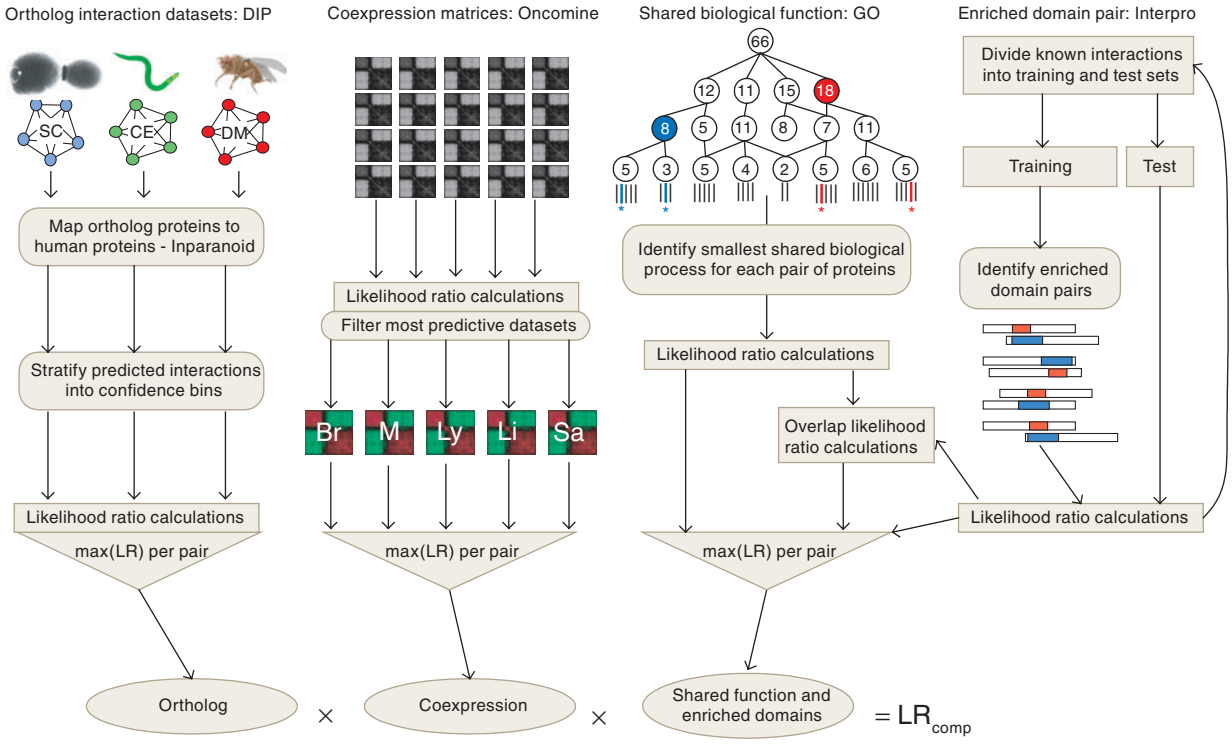

Figure 2 Data integration in a semi-naïve Bayes model to predict human protein-protein interactions. Four distinct types of evidence, listed at the top of the figure with the source of the data, were found to predict human protein-protein interactions. Step 1 was to calculate likelihood ratios for groups of predictions based each data type (**Fig. 1**). Step 2 was to identify data types that contain redundant information and then calculate likelihood ratios for predictions with both data types (shared biological function and enriched domain pair). Step 3 was to identify the maximum likelihood ratio, max(LR), from each evidence type. Step 4 was to multiplicatively combine likelihood ratios in a naive Bayes model to generate composite likelihood ratios (LR_{comp}). Gene expression data sets: Br, Breast; Li, Liver; Ly, Lymphoma; M, Multi-cancer; Sa, Sarcoma.

from a single SC interaction (n = 1,425; LR = 490.2), a medium confidence bin, in which more than one but less than 28 human interactions were predicted per SC interaction (n = 9,008; LR = 67.0), and a low-confidence bin, in which more than 28 human interactions were predicted per SC interaction (n = 6,725; LR = 9.8). The DM data were divided into two bins: a medium-high confidence bin, in which the DM interactions had a confidence score greater than 0.55 (n = 3,088; LR = 324.4), and a low confidence bin, in which the confidence score was less than 0.55 (n = 14,030; LR = 9.2). The decision tree algorithm did not stratify the interactions predicted from the CE data. We also created a separate bin for 201 predicted human protein interactions that were predicted by two model organism data sets (n = 201; LR = 1664.8; **Fig. 1a**).

Gene expression

Although it is well known that interacting proteins are often coexpressed, it is unclear if data on coexpression can be used to predict human protein-protein interactions. We identified coexpressed genes from 65 genome-wide gene expression data sets present in the Oncomine Cancer Microarray Database[3,21]. The 65 data sets consisted of more than 5,000 diverse microarray profiles representing many tissue types, differentiation states and cellular compositions, thus covering a broad spectrum of gene expression (**Supplementary Table 2**). For each data set, we calculated Pearson correlations for each pair of genes and then grouped gene-pairs into 19 correlation bins. Likelihood ratios were then calculated for each correlation bin in each data set. Nearly all data sets had a weak positive association between coexpression and protein interactions. In five selected data sets, each of which profiled more than 80 human tissue samples, the likelihood ratios were considerably stronger and increased consistently with increasing coexpression (**Fig. 1b** and **Supplementary Table 3**).

Shared biological function

Two proteins that function in the same biological process (for example, the cell cycle) should be more likely to interact than two proteins that do not. Furthermore, proteins functioning in small, specific biological processes should be more likely to interact than proteins functioning in large, general processes (for example, mitotic spindle checkpoint versus cell proliferation). We downloaded biological process annotations from the Gene Ontology Consortium[4] and compressed the hierarchy to derive 94,045 assignments of 9,345 proteins to one or more of 1,887 biological processes. Next, as a measure of functional similarity, we identified the smallest shared biological process (SSBP) for each pair of annotated proteins, binned protein pairs by this measure, and then generated likelihood ratios for each bin by testing against the GSP and GSN. As expected, protein pairs with shared biological function annotations were more likely to interact than those without shared annotations, and protein pairs that shared small, specific functional annotations were more likely to interact than pairs sharing large, general annotations (**Fig. 1c** and **Supplementary Table 4**). For example, protein pairs that shared a biological function annotation with more than 10 but fewer than 50 total proteins were far more likely to overlap with the GSP (6.9%)

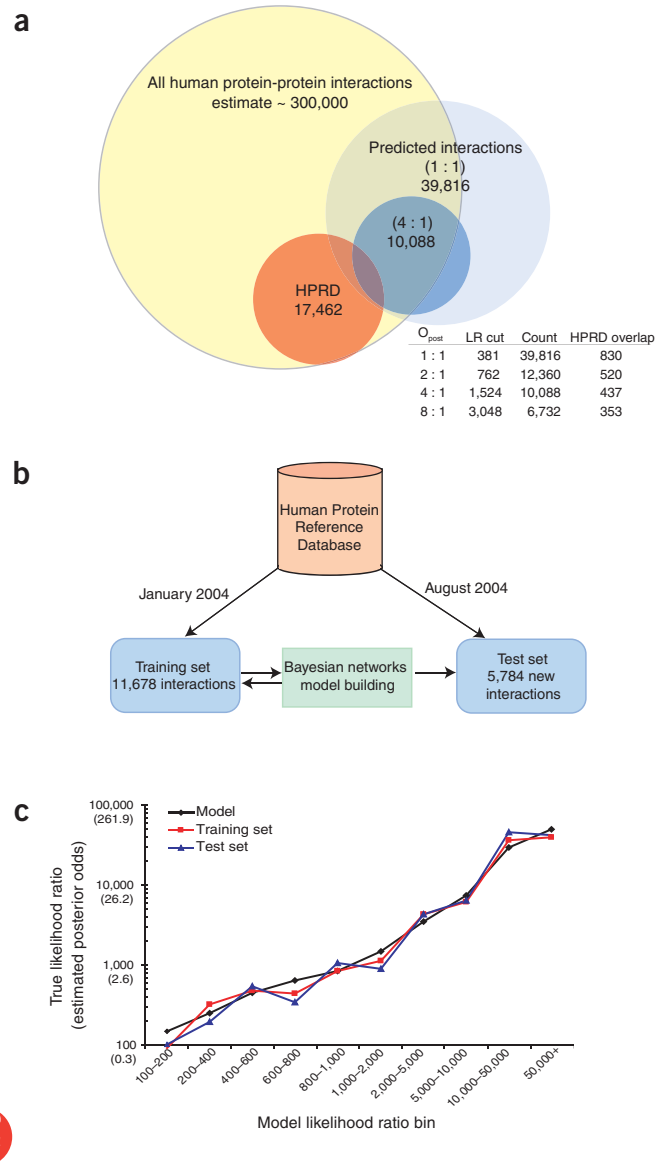

Figure 3 Characterization and performance analysis of the predicted interactome. (**a**) The overlap of the predicted and known components of the human protein interactome. The total number of human protein interactions was estimated by assuming that the actual overlap between the known and predicted components of the interactome is equal to the overlap that would be expected by chance when randomly selecting the two equivalently sized subsets of the interactome. Inset: O_{post}, posterior odds ratio. For example, 8:1 means that there are 8-to-1 odds that a pair of proteins interacts given the predictive data sets. LR_{cut}, likelihood ratio cutoffs corresponding to selected odds ratios. Count, the number of protein-protein interactions predicted at the selected LR_{cut} values. (**b**) A schematic detailing how literature-referenced protein-protein interactions from the Human Protein Reference Database were used for training and testing a Bayes model for the prediction of protein-protein interactions. (**c**) Performance of the Bayes model, when applied to the training set, and an independent test set of known protein-protein interactions. Likelihood ratios based on the model were ascribed to protein pairs in the training and test sets, the training and test set protein pairs were binned based on the model likelihood ratios, and then actual likelihood ratios were computed for each bin. The expected likelihood ratio was compared to the actual likelihood ratios.

than with the GSN (0.037%), generating a moderately strong predictor of protein interactions (LR = 188.8).

Protein domains

Because protein interactions involve physical associations between protein domains, it has been proposed that novel protein interactions may be predicted by identifying pairs of domains enriched among known interacting proteins[9]. To test this logic in the context of our GSP and GSN sets, we downloaded the Interpro database, which consisted of 19,438 assignments of 1,352 protein domains and families to one or more of 9,779 proteins. To quantify the co-occurrence of particular domain pairs among interacting proteins, we devised the domain enrichment ratio (D). Because the identification of predictive domain pairs requires a large set of known protein interactions, we divided the GSP into thirds, and used two thirds of the GSP to define enriched domain pairs and the remaining third to test the ability of the enriched domain pairs to predict new interactions. We repeated this process three times and combined the results. We found that the degree of domain enrichment in two thirds of the GSP is strongly associated with the likelihood of interaction in the remaining third (**Fig. 1d** and **Supplementary Table 5**). Pairs of domains with large domain enrichment ratios (D > 10) were strongly predictive of protein interactions (LR = 322.4), whereas progressively smaller D values were less strongly associated with protein interactions (3 < D < 10; LR = 35.7; 2 < D 3, LR = 5.8). An example of an enriched domain pair is Hedgehog signaling domain and Sterol-sensing 5TM box. Five of 7,671 interactions (0.065%) were between a protein with the hedgehog domain and a protein with the sterol domain. Given the respective frequencies of these domains, we would expect just 0.00033% of interactions to be between proteins with these domains (D = 199.9). We suspect that many of the enriched domain pairs represent physically interacting domains; however, it is also possible that the co-occurrence is due to indirect interactions. A complete list of enriched domain pairs is available as **Supplementary Table 6** online.

Integrative analysis: naive Bayes classifier

Thus far, we have predicted a large number of protein-protein interactions with low confidence and a small number of interactions with high confidence. To predict a larger number of protein-protein interactions with high confidence, we probabilistically combined the evidence sources in a naive Bayes model, which multiplicatively combines the data set–specific likelihood ratios. This multiplicative nature requires that the predictive data sets be conditionally independent or nonredundant. We found that the four types of evidence were essentially independent, except for the shared biological function and enriched domain pair evidence types, which were found to contain redundant information (**Supplementary Methods** and **Supplementary Table 8** online). To avoid bias, we computed separate likelihood ratios for protein pairs having predictive evidence from both data sources. Using a decision tree algorithm, we created five bins: a high confidence bin containing protein pairs with strong domain pair enrichment (D > 10) and some degree of functional similarity (SSBP < 1,000; n = 4253; LR = 19,381), a medium-high confidence bin containing protein pairs with moderate domain pair enrichment (3 < D < 10) and some degree of similarity in biological function (SSBP < 1000; n = 16,348; LR = 495.6), another medium-high confidence bin containing protein pairs with weak domain pair enrichment (2 < D < 3) but strong functional similarity (SSBP < 100; n = 4,117; LR = 407.2), and finally, two low confidence bins (**Fig. 1d** and **Supplementary Table 8**).

The predictive data sets were combined in a naive Bayes model, which was applied to all protein pairs, automatically combining available evidence sources to derive composite likelihood ratios (LR_{comp}; **Fig. 2**). This resulted in the prediction of 39,816 interactions at a likelihood

ratio cutoff (LR_{cut}) of 381 (O_{post} = 1) and 10,088 interactions at LR_{cut} of 1,521 (O_{post} = 4). The full list of predicted interactions is available (**Supplementary Table 9** online). **Figure 3a** relates predicted interactions at various confidence levels with the GSP interactions and the estimated superset of all human protein-protein interactions. The result of nearly 40,000 predicted interactions with a false positive rate of 50% and more than 10,000 predicted interactions with a false positive rate of just 20% is comparable or superior to the results of high-throughput experimental approaches in model organisms[13–18]. To examine the validity of this model, we binned predicted interactions by LR_{comp} and assessed the true likelihood ratios for each bin, based on the intersection with the GSP and GSN (**Fig. 3b** and **Supplementary Table 10** online). As anticipated, the true likelihood ratios closely parallel the LR_{comp} measure provided by the model, confirming that the multiplicative nature of the model does not overestimate the likelihood of interaction in the training set. Next, we assessed the model on an independent test set of 5,784 known interactions that were queried from the Human Protein Reference Database after the model was built. As shown in **Figure 3c**, the model performs similarly on the training and test sets, suggesting that our model provides a valid measure of the odds that two proteins interact. **Figure 4a** is a global view of 10,088 high-confidence predicted interactions (LR_{comp} > 1526; O_{post} > 4) among 3,039 proteins.

Network analysis and experimental validation

To explore the complex interaction circuitry among human proteins at the level of a single protein or pathway, we created a public bioinformatics resource, Human Interactome Map or HiMAP (http://www.himap.org). Using HiMAP, we investigated two specific areas of the interactome map and identified intriguing predicted interactions, which we later confirmed experimentally. First, we sought to identify previously unknown components of the mitotic spindle checkpoint, as this pathway is important in aneuploidy and cancer[22,23]. Beginning with several well characterized members of the mitotic spindle checkpoint (CDC20, BUB1 and BUB3, among others), we found many novel predicted interactions, including one between BUB3 and ZNF207, an uncharacterized zinc finger protein (**Fig. 4b**). Next, we seeded HiMAP with a protein of interest, RSU1, which is a potential tumor suppressor, whose gene is down-regulated in prostate cancer[24–26]. Because the mechanism by which *RSU1* exerts its tumor suppressor effects remain to be elucidated, we sought to implicate this protein in a specific pathway. HiMAP analysis revealed that RSU1 was predicted to interact with LIMS1 which forms a ternary complex with ILK (integrin-linked kinase) and NCK2 and has been shown to colocalize with integrins[27] (**Fig. 4c**).

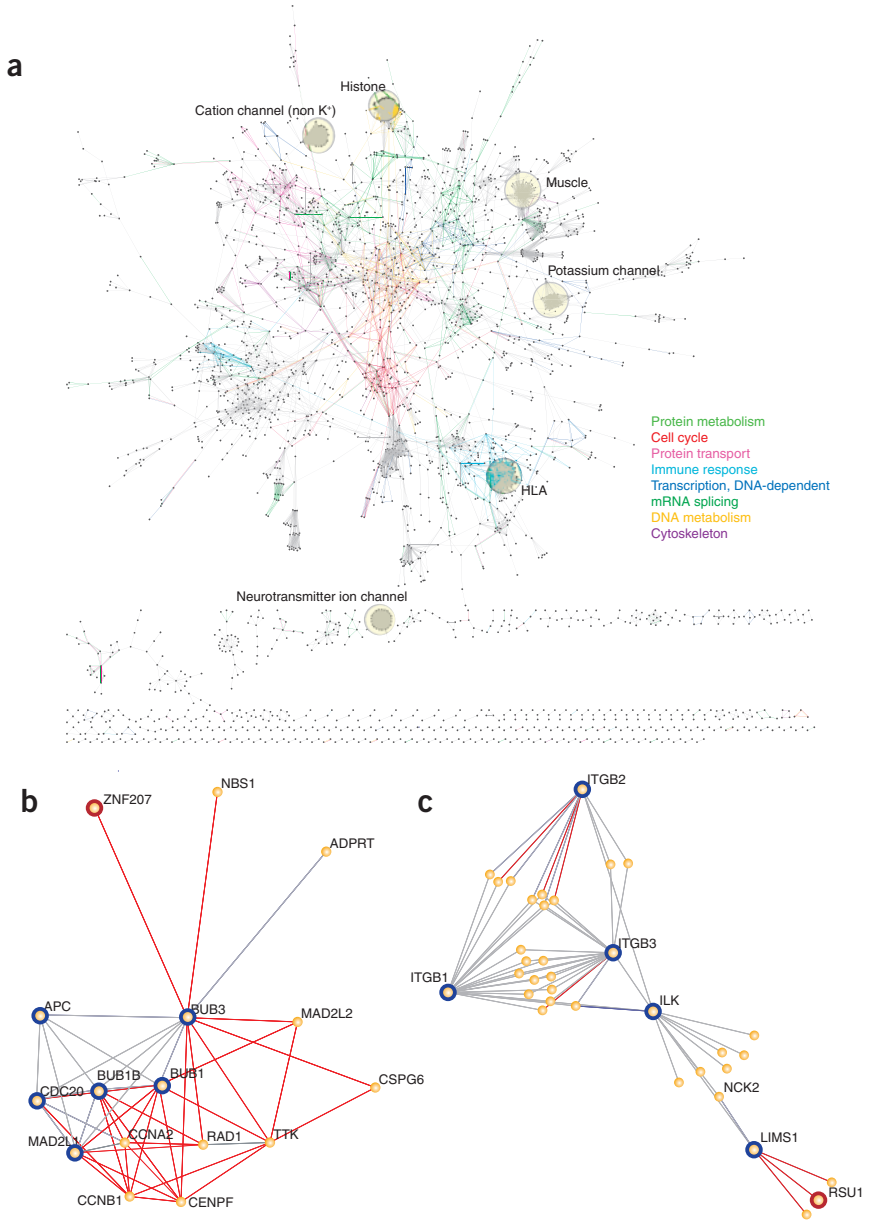

Figure 4 Global and focused views of the predicted human interactome. (**a**) We predicted 10,088 protein-protein interactions (edges) among 3,039 proteins (nodes) with high confidence (LR > 1,524; O_{post} > 4). Interacting proteins that function in one of eight selected biological processes are colored as indicated, and 'interaction cliques' among highly related proteins are labeled and highlighted with pale yellow circles. HLA, human leukocyte antigens. An interaction map with gene names is available in **Supplementary Figure 1** online. (**b**,**c**) Zoomed in views of the predicted interactome generated using the online resource, HiMAP. Known interactions are shown in gray and predicted interactions in red. Well-characterized members of the respective pathways are designated with dark blue circles and experimentally confirmed new members with red circles. ZNF207 was predicted and experimentally confirmed to interact with BUB3, potentially implicating this uncharacterized zinc finger transcription factor in the mitotic checkpoint pathway (**b**). RSU1 was predicted and experimentally confirmed to interact with LIMS1, an integrin-mediated signaling adaptor protein, downstream of integrins and integrin-linked kinase, suggesting a pathway through which RSU1 may exert its documented tumor suppressor effects (**c**).

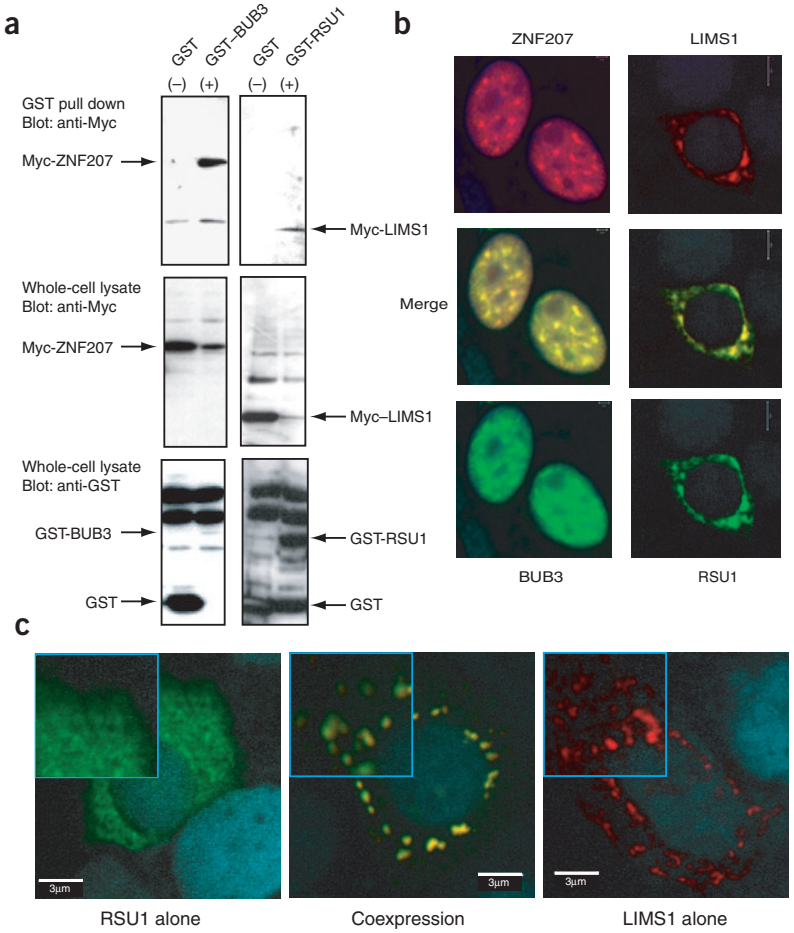

Figure 5 Experimental confirmation of two predicted interactions implicates uncharacterized proteins into specific pathways. (**a**) Coimmunoprecipitation. Mammalian 293T cells were transfected with GST-tagged fusion proteins (or GST alone) and Myc-tagged fusion proteins of predicted interacting partners. Cell lysates were precipitated with glutathione-Sepharaose and immunoblotted with anti-Myc (upper panels). The total cell lysates before precipitation were also directly immunoblotted to confirm expression of Myc-tagged proteins (middle panels) and GST-tagged proteins or GST (lower panels). The expected locations of the tagged fusion proteins and control GST are indicated. (**b**) Colocalization of interacting proteins. MCF-7 cells were cotransfected with the indicated GST-bait protein expression vector (green channel, bottom) and Myc-prey protein expression vector (red channel, top). The merged images are in the middle. Fusion proteins were detected using mouse anti-Myc and rabbit anti-GST, and signals were visualized using Alexa flour dyes (Alexa 488 and Alexa 555, respectively). DAPI (blue channel) was used for visualization of the nucleus. (**c**) Relocalization of RSU1 when coexpressed with LIMS1. GST-tagged RSU1 expression vector and Myc-tagged LIMS1 were expressed individually (right and left, respectively) or coexpressed (middle) and visualized as above. The insets highlight the altered localization of RSU1 when coexpressed with LIMS1. The first inset is a zoomed image demonstrating diffuse staining at high power.

Using coimmunoprecipitation assays, we confirmed the predicted interactions between BUB3 and ZNF207 and between RSU1 and LIMS1 (**Fig. 5a**). To validate the biological relevance of these interactions, we also demonstrated colocalization by confocal microscopy (**Fig. 5b**). The BUB3-ZNF207 interaction is notable because it represents a new link between the well-established mitotic spindle checkpoint pathway and an uncharacterized zinc finger transcription factor that, to date, had not been expressed or characterized at the protein level[28]. The RSU1-LIMS1 interaction is notable because it implicates a potential tumor suppressor protein (RSU1) in the integrin signaling pathway. *RSU1* was originally identified in a screen for genes that could inhibit Ras-mediated signaling[29] and was later shown to inhibit anchorage-independent growth of MCF7 breast cancer cells[30]. Notably, we have observed that *RSU1* is consistently downregulated in prostate cancer[3,24,26], further suggesting a tumor suppressor role for this gene; however, to date, the mechanism by which *RSU1* exerts its effects has not been elucidated. Here our model predicted, and we experimentally confirmed that RSU1 interacts with LIMS1, an adaptor protein involved in integrin signaling at focal adhesions[31]. Furthermore, we found that when RSU1 was expressed alone, it had diffuse cytoplasmic localization, but when it was coexpressed with LIMS1, it had discrete punctate cytoplasmic localization (**Fig. 5c**). These results are consistent with the localization of LIMS1 to focal adhesions and demonstrate that LIMS1 recruits RSU1 to focal adhesions, suggesting that RSU1 may exert its effects through LIMS1 and the integrin signaling pathway.

Next we explored the utility of the known and predicted interactome networks for interpreting cancer gene expression data. We generated protein interaction subnetworks for genes overexpressed in pancreatic adenocarcinoma, multiple myeloma and renal cell carcinoma, as defined by Oncomine[3]. In pancreatic adenocarcinoma, we found an oncogenic tyrosine kinase subnetwork involving ERBB2, MUC1, SHC1 and EPH2A and an invasion signaling subnetwork involving NET1, RhoA, RhoC and RAC (**Supplementary Fig. 2** online). RhoA, RhoC and RAC are all small G-protein signaling molecules with known roles in cell migration and metastasis[32–34], whereas NET1 is a guanine nucleotide exchange factor, which activates RhoA and RhoC[35] and has transforming ability on its own[36]. Perhaps overexpression of this subnetwork is responsible for the invasive properties of pancreatic adenocarinoma. In multiple myeloma, we identified an activated oncogenic signaling subnetwork involving H-RAS, RAF1, BAG1 and PAK1 (**Supplementary Fig. 3** online). H-RAS is a small GTPase that undergoes activating mutations in several cancers[37], whereas RAF1 is a downstream mediator of H-RAS signaling that on its own can induce RAS-like tumorigencity[38]. Also, BAG1 and PAK1 are capable of binding to and activating RAF1[39,40]. Finally, in the clear cell variant of renal cell carcinoma, we identified an oncogenic subnetwork of activated proteins including VEGF, KDR, PDGFB, PDGFRB, SHC, MAPK1, CSF1R, FYN and LYN (**Supplementary Fig. 4** online). This subnetwork is unique because it includes two overexpressed receptor ligand pairs, suggesting active autocrine signaling loops.

In summary, we have integrated disparate genomic and proteomic data sources to develop a model for predicting human protein-protein interactions on a global scale. Whereas previous studies have attempted

to infer protein interactions from genomic data sources[6,41], we have built a robust, comprehensive and presumably more accurate predictive model by probabilistically combining several independent data sources and calculating reliable confidence measures by rigorous testing against large sets of gold standard positive and negative interactions. Furthermore, we proved the validity of our model on an independent test set of known interactions and experimentally confirmed two predicted interactions, expanding the mitotic spindle checkpoint pathway and implicating a potential tumor suppressor gene in the integrin signaling pathway. We have also built HiMAP for analyzing the known and predicted components of the human interactome. Lastly, we demonstrated the utility of the human interactome for interpreting genome-wide gene expression data in complex human diseases such as cancer. We anticipate that the predictive model will grow in size and accuracy as the GSP set is expanded and as new predictive evidence sources become available.

METHODS

Detailed methods. Additional methods are available in **Supplementary Methods** online.

Gold standard interactions. The GSP interaction set was downloaded from the HPRD (http://www.hprd.org). In January, 2004, 11,678 interactions among 5,505 proteins, and the literature references were downloaded from HPRD. Later, in August, 2004, 5,784 new interactions were downloaded, which were used as the independent test set. None of the test set interactions were part of the training set. The GSN interaction set was defined as all protein pairs in which one protein was assigned the plasma membrane cellular component (1,426 proteins) and the other the nuclear cellular component (2,253), as assigned by Gene Ontology Consortium. Twenty-nine proteins that were assigned to both components were removed. In total, 3,106,928 unique pairs were identified.

Integrated data sets. Publicly available model organism protein interaction data sets were downloaded from the DIP (http://dip.doe-mbi.ucla.edu/dip/Download.cgi). Pairwise ortholog map files were downloaded from the Inparanoid database (http://inparanoid.cgb.ki.se/). Logical bins based on several parameters associated with the predicted interactions were defined using the J48 decision tree algorithm as implemented in the Weka software package[20] (http//www.cs.waikato.ac.nz/ml/weka).

To identify genes that are coexpressed, publicly available microarray data were collected from the Oncomine Cancer Microarray Database (http://www.oncomine.org). Sixty-five data sets were available and analyzed independently. Pearson correlations were computed between all pairs of genes with values present in 50% of the profiled samples, and then gene pairs were grouped into 19 correlation bins of increasing coexpression. Five data sets were selected for the final analysis: a multi-cancer data set profiling of 174 cancer samples of 11 tissue types[42], a breast cancer data set profiling 117 breast tumors[43], a liver data set profiling 197 normal and cancerous liver samples[44], a lymphoma data set profiling 293 lymphoma samples[45] and a soft-tissue data set profiling 81 melanomas and soft-tissue tumors[46]. Biological process annotations were downloaded from the Gene Ontology Consortium[4] and compressed the hierarchy to derive 94,045 assignments of 9,345 proteins to one or more of 1,887 biological processes. The SSBP per pair of proteins was defined. Protein domain and family assignments were downloaded from the Interpro database. In total, 19,438 assignments of 1,352 protein domains and families to one or more of 9,779 proteins were queried. Domain pair enrichment was assessed with the domain enrichment ratio (D), which is calculated as the probability (Pr) of observing a pair of domains in a set of known interacting proteins divided by the product of the probabilities of observing each domain pair independently:

$$D = \Pr(d_i{:}d_j \mid GSP) \div (\Pr(d_i \mid GSP) \times \Pr(d_j \mid GSP))$$
$$d_i{:}d_j \geq 3,$$

where d_i and d_j are two protein domains, $d_i{:}d_j$ is a protein-protein interaction in which one protein has d_i and one has d_j, and GSP is a gold standard positive set of known interactions. We also ensured that a minimum representation of at least three interactions was present in the GSP set.

Naive Bayes classifier approach. Following a derivation of Bayes rule, the posterior odds of interaction (O_{post}) can be calculated as the product of the prior odds of interaction (O_{prior}) and the likelihood ratio, L (f1). The prior odds being the chance of choosing a pair of interacting proteins from all protein pairs and the likelihood ratio being the probability of observing the values in the predictive data sets given that a pair of proteins interacts divided by the probability of observing the values given that the pair does not interact (f2). In the special case in which the predictive data sets are conditionally independent or nonredundant, the likelihood ratio can be calculated as the product of individual data-set likelihood ratios (f3). Formally stated:

$$(f1)\ O_{post} = O_{prior} \times L$$

$$(f2)\ L = \Pr(f_1 \ldots f_n \mid GSP) \div \Pr(f_1 \ldots f_n \mid GSN)$$

$$(f3)\ L = \prod_{i=1}^{i=n} \Pr(f_i \mid GSP) \div \Pr(f_i \mid GSN),$$

where L is the likelihood ratio, f is a protein pair's value in data sets *i*, GSP is a gold standard positive set of known interactions, and GSN is a gold standard negative set of protein pairs that do not interact.
The prior odds of interaction were defined as:

$$O_{prior} = P(pos) / P(neg),$$

where P(pos) is the probability of finding an interacting pair of proteins among all pairs of proteins, and P(neg) is the probability of finding a non-interacting pair. The prior odds were estimated by examining the average number of interactions per protein for which all known interactions were identified in the literature. Among 2,987 proteins, 11,678 distinct interactions existed, thus the probability that two randomly selected proteins interact was calculated to be 1 in 382. The posterior odds or the odds that two proteins interact given new predictive evidence were defined as:

$$O_{posterior} = P(pos \mid f_1 \ldots f_n) \div P(neg \mid f_1 \ldots f_n)$$

Where f_i is a protein pair's value in data set i. The likelihood ratio:

$$L = (f_1 \ldots f_n \mid pos) \div P(f_1 \ldots f_n \mid neg)$$

Relates the prior odds and the posterior odds as defined by a derivation of Bayes rule:

$$O_{posterior} = O_{prior} \times L(f_1 \ldots f_n)$$

When the evidence types integrated are independent (or non-redundant), the likelihood ratio can be calculated simply as the product of individual likelihood ratios from the respective evidence types. This is known as a Naive Bayes Network:

$$L(f_1 \ldots f_n) = \prod_{i=1}^{i=n} (f_i \mid pos) \div P(f_i \mid neg)$$

Because the Shared Biological Function and Domain Enrichment evidence types were found to be semiredundant, they were analyzed together, so that only one likelihood ratio was submitted from these two data sources. Also, because coexpression in multiple data sets was found to be redundant information, only the largest coexpression likelihood ratio was submitted to the model.

Interaction network graphing. The HiMAP web application (http://www.himap.org) was developed to dynamically visualize and explore a database of protein interactions. The application was written in Java and uses an Oracle 9i database. The r-PolyLog energy model was implemented to lay out interaction networks, and the networks are displayed with scalable vector graphics (SVG). We created **Figure 4a** with Cytoscape[47] (http//www.cytoscape.org).

Open reading frame cloning and coaffinity purification experiments. Full-length, sequence-verified mammalian gene collection (MGC) clones (Open Biosystems) were obtained for the indicated proteins and Gateway-cloned essentially as described[48]. Clones and primer sequences used for amplification are available in **Supplementary Table 11** online. ORFs were amplified using adapter PCR and cloned into the Gateway vector pDNR221 (Invitrogen). Entry clones of bait proteins were subcloned into pDEST-27, which contains the glutathione *S*-transferase (GST) epitope tag upstream of a Gateway recombination site (gift of A. Swaroop, University of Michigan Medical School). Entry clones of prey proteins were subcloned into pCMV-Myc-DEST, which contains a Myc epitope tag upstream of the Gateway recombination site (Kind gift of J. W. Harper, Harvard Medical School).

For co-affinity purification experiments, 1 μg of each plasmid was transfected into 293T cells using Fugene 6 reagent according to the manufacturer's instructions (Roche). For GST control plasmids, 0.5 μg of each plasmid was used per transfection. Cells were cultured for 2 d in RPM1 medium with 10% fetal bovine serum (Invitrogen) and lysed in 0.5% NP-40 buffer (20 mM Tris-HCl (pH 8.0), 100 mM NaCl, 1 mM EDTA and complete protease inhibitor cocktail (Roche)). Lysates were cleared by centrifugation at 14,000*g* before precipitation of protein complexes using glutathione-Sepharose beads. Beads were washed three times with lysis buffer, and purified complexes and control lysate samples were separated on acrylamide gels (Bio-Rad). Myc- and GST-tagged proteins were detected using standard immunoblotting techniques. Primary antibodies used were mouse monoclonal anti-Myc (clone 9E10, Cell Signaling Technology) and rabbit polyclonal anti-GST (Sigma).

Immunofluorescence and confocal microscopy. The breast carcinoma cell line MCF-7 was grown on chamber slides (Lab-Tek) overnight in RPMI medium and transfected with Myc-tagged or GST-tagged clones alone in individual chambers or cotransfected with both clones for 36 h. Cells were washed and fixed using chilled methanol. The slides were then blocked in PBS-T (phosphate-buffered saline with Tween-20 (0.01%) with 5% normal donkey serum for 1 h. A mixture of rabbit anti-GST (Sigma) and mouse anti-Myc (Cell Signaling Technology) was added to the slides at 1:500 and 1:1,000 dilutions, respectively, and incubated for 1 h at 15–25 °C. Slides were then incubated with secondary antibodies (anti-rabbit Alexa 488 and anti-mouse Alexa 555 (Molecular Probes) at 1:1,000 dilution) for 1 h. After washing the slides with PBS-T and PBS, the slides were mounted using VECTASHIELD mounting medium containing DAPI (4',6-diamidino-2-phenylindole; Vector Laboratories). Confocal images were taken with a Zeiss LSM510 META (Carl Zeiss) imaging system using ultraviolet, argon and helium neon 1 light source. The triple color images were exported as TIFF images.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank R. Varambally for database assistance, D. Gibbs for hardware support, and the Institute of Bioinformatics for making the Human Protein Reference Database available. This work was funded by pilot funds from the Dean's Office, Department of Pathology, Cancer Center Support Grant P30 CA46592, and the Bioinformatics Program. D.R.R. and S.A.T. are fellows of the Medical Scientist Training Program, D.R.R. was funded by the Cancer Biology Training Program and A.M.C. is a Pew Scholar. A.P. is chief scientific advisor to the Institute of Bioinformatics. The terms of this arrangement are being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
2. Mulder, N.J. *et al.* The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318 (2003).
3. Rhodes, D.R. *et al.* ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1–6 (2004).
4. Harris, M.A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
5. Yu, H. *et al.* Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).
6. Huang, T.W. et al. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics* **20**, 3273–3276 (2004).
7. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
8. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29**, 482–486 (2001).
9. Ng, S.K., Zhang, Z. & Tan, S.H. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* **19**, 923–929 (2003).
10. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
11. Marcotte, E.M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
12. 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).
13. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
14. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
15. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
16. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
17. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
18. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
19. Remm, M., Storm, C.E. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
20. Witten, I.H. & Frank, E. Data Mining: Practical machine learning tools with Java implementations. (Morgan Kaufmann, San Francisco, 2000).
21. Rhodes, D.R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. USA* **101**, 9309–9314 (2004).
22. Cahill, D.P. *et al.* Mutations of mitotic checkpoint genes in human cancers. *Nature* **392**, 300–303 (1998).
23. Bharadwaj, R. & Yu, H. The spindle checkpoint, aneuploidy, and cancer. *Oncogene* **23**, 2016–2027 (2004).
24. Dhanasekaran, S.M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826 (2001).
25. Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. & Chinnaiyan, A.M. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* **62**, 4427–4433 (2002).
26. Welsh, J.B. *et al.* Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.* **61**, 5974–5978 (2001).
27. Tu, Y., Li, F., Goicoechea, S. & Wu, C. The LIM-only protein PINCH directly interacts with integrin-linked kinase and is recruited to integrin-rich sites in spreading cells. *Mol. Cell. Biol.* **19**, 2425–2434 (1999).
28. Pahl, P.M. *et al.* ZNF207, a ubiquitously expressed zinc finger gene on chromosome 6p21.3. *Genomics* **53**, 410–412 (1998).
29. Cutler, M.L., Bassin, R.H., Zanoni, L. & Talbot, N. Isolation of rsp-1, a novel cDNA capable of suppressing v-Ras transformation. *Mol. Cell. Biol.* **12**, 3750–3756 (1992).
30. Vasaturo, F., Dougherty, G.W. & Cutler, M.L. Ectopic expression of Rsu-1 results in elevation of p21CIP and inhibits anchorage-independent growth of MCF7 breast cancer cells. *Breast Cancer Res. Treat.* **61**, 69–78 (2000).
31. Fukuda, T., Chen, K., Shi, X. & Wu, C. PINCH-1 is an obligate partner of integrin-linked kinase (ILK) functioning in cell shape modulation, motility, and survival. *J. Biol. Chem.* **278**, 51324–51333 (2003).
32. Ikoma, T. *et al.* A definitive role of RhoC in metastasis of orthotopic lung cancer in mice. *Clin. Cancer Res.* **10**, 1192–1200 (2004).
33. Schroeder, J.A., Thompson, M.C., Gardner, M.M. & Gendler, S.J. Transgenic MUC1 interacts with epidermal growth factor receptor and correlates with mitogen-activated protein kinase activation in the mouse mammary gland. *J. Biol. Chem.* **276**, 13057–13064 (2001).
34. Michiels, F., Habets, G.G., Stam, J.C., van der Kammen, R.A. & Collard, J.G. A role for Rac in Tiam1-induced membrane ruffling and invasion. *Nature* **375**, 338–340 (1995).
35. Alberts, A.S. & Treisman, R. Activation of RhoA and SAPK/JNK signalling pathways by the RhoA-specific exchange factor mNET1. *EMBO J.* **17**, 4075–4085 (1998).
36. Chan, A.M., Takai, S., Yamada, K. & Miki, T. Isolation of a novel oncogene, NET1, from neuroepithelioma cells by expression cDNA cloning. *Oncogene* **12**, 1259–1266 (1996).
37. Cerutti, P., Hussain, P., Pourzand, C. & Aguilar, F. Mutagenesis of the H-ras protooncogene and the p53 tumor suppressor gene. *Cancer Res.* **54**, 1934s–1938s (1994).
38. Khosravi-Far, R. *et al.* Oncogenic Ras activation of Raf/mitogen-activated protein kinase-independent pathways is sufficient to cause tumorigenic transformation. *Mol. Cell. Biol.* **16**, 3923–3933 (1996).
39. Wang, H.G., Takayama, S., Rapp, U.R. & Reed, J.C. Bcl-2 interacting protein, BAG-1, binds to and activates the kinase Raf-1. *Proc. Natl. Acad. Sci. USA* **93**, 7063–7068 (1996).

40. Zang, M., Hayne, C. & Luo, Z. Interaction between active Pak1 and Raf-1 is necessary for phosphorylation and activation of Raf-1. *J Biol. Chem.* **277**, 4395–4405 (2002).
41. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
42. Su, A.I. *et al.* Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* **61**, 7388–7393 (2001).
43. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
44. Chen, X. *et al.* Gene expression patterns in human liver cancers. *Mol. Biol. Cell* **13**, 1929–1939 (2002).
45. Rosenwald, A. *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346**, 1937–1947 (2002).
46. Segal, N.H. *et al.* Classification of clear-cell sarcoma as a subtype of melanoma by genomic profiling. *J. Clin. Oncol.* **21**, 1775–1781 (2003).
47. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
48. Tewari, M. *et al.* Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-β signaling network. *Mol. Cell* **13**, 469–482 (2004).