In silico prediction of physical protein interactions and characterization of interactome orphans

Max Kotlyar[1], Chiara Pastrello[1,2], Flavia Pivetta[2], Alessandra Lo Sardo[2], Christian Cumbaa[1], Han Li[3], Taline Naranian[3], Yun Niu[1,4], Zhiyong Ding[5], Fatemeh Vafaee[1,6], Fiona Broackes-Carter[1], Julia Petschnigg[7], Gordon B Mills[5], Andrea Jurisicova[3], Igor Stagljar[7], Roberta Maestro[2] & Igor Jurisica[1,8–10]

Protein-protein interactions (PPIs) are useful for understanding signaling cascades, predicting protein function, associating proteins with disease and fathoming drug mechanism of action. Currently, only ~10% of human PPIs may be known, and about one-third of human proteins have no known interactions. We introduce FpClass, a data mining–based method for proteome-wide PPI prediction. At an estimated false discovery rate of 60%, we predicted 250,498 PPIs among 10,531 human proteins; 10,647 PPIs involved 1,089 proteins without known interactions. We experimentally tested 233 high- and medium-confidence predictions and validated 137 interactions, including seven novel putative interactors of the tumor suppressor p53. Compared to previous PPI prediction methods, FpClass achieved better agreement with experimentally detected PPIs. We provide an online database of annotated PPI predictions (http://ophid.utoronto.ca/fpclass/) and the prediction software (http://www.cs.utoronto.ca/~juris/data/fpclass/).

Determining a comprehensive human PPI network is a major goal in the post-genomic era[1]. Over the last decade, about 10,000 new PPIs have been reported each year, enlarging the known human interactome from 16,276 to 114,906 PPIs[2]. However, the known PPI network is far from complete. Estimates of the total number of human PPIs range from 130,000 to over 600,000 (refs. 3–5) and may not account for transient interactions[4]. More importantly, the known interaction network is missing about one-third of human proteins (as they have no known interactions), and another third have fewer than five known interactions. Consequently, any network-based analysis, such as prediction of protein function or disease roles, is limited to a fraction of the proteome. Although some proteins may actually participate in only a few interactions, in many cases, research bias[6] and limitations of biological assays[7–9] are likely responsible for the sparse interactome. Computational PPI prediction methods can address this challenge by identifying candidate missing interactions or prioritizing targets for high-throughput screens, which may reduce the cost of mapping the interactome[10].

We present an *in silico* method, FpClass, to predict high-confidence PPIs proteome-wide, including proteins with few (low-degree proteins) or no known partners (orphans). Prediction for such proteins is challenging because these proteins are seldom studied and are usually poorly annotated. We introduce three strategies for ensuring high-confidence predictions, even for proteins with limited interaction evidence, and for achieving low false discovery rate (FDR). First, FpClass identifies sets of protein features (for example, domains or post-translational modifications (PTMs)) that may act cooperatively, making an interaction more likely if a protein possesses the entire set of features rather than individual ones. FpClass uses such sets as additional predictive features. Second, whereas other prediction methods[11,12] search for pairs of compatible domains that mediate interactions, FpClass searches for diverse types of compatible features. For example, it estimates the likelihood of an interaction by taking into account pairs of apparently unrelated features that tend to co-occur in interacting partners (for example, a specific domain in one partner and the localization of the second partner). Third, it searches for incompatible pairs of features that may reduce the chances of interaction.

PPI prediction proceeds in four steps: (i) defining a training set of interacting and non-interacting protein pairs annotated with diverse evidence (including protein features and network topology), (ii) identifying frequent feature sets, (iii) determining the probability of interaction given particular evidence, and (iv) predicting interactions by combining probabilities from available evidence (**Fig. 1**). A number of different data sets were used for training and testing FpClass. Each data set comprised PPIs detected by biological assays (positive cases) as well as random protein pairs representing non-interactions (negative cases). The largest training set, formed by the union of the smaller sets,

[1]Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada. [2]Centro Riferimento Oncologico, Istituto Nazionale Tumori, Aviano, Italy. [3]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [4]Nanjing University of Aeronautics and Astronautics, Nanjing, China. [5]Department of Systems Biology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [6]Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia. [7]Donnelly Centre, Departments of Molecular Genetics and Biochemistry, University of Toronto, Toronto, Ontario, Canada. [8]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [9]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [10]TECHNA Institute for the Advancement of Technology for Health, Toronto, Ontario, Canada. Correspondence should be addressed to I.J. (juris@ai.utoronto.ca).

RECEIVED 10 DECEMBER 2013; ACCEPTED 14 AUGUST 2014; PUBLISHED ONLINE 17 NOVEMBER 2014; CORRECTED ONLINE 10 DECEMBER 2014 (DETAILS ONLINE);
DOI:10.1038/NMETH.3178

comprised 20,454 PPIs detected in at least two independent experiments (**Supplementary Data 1**) and 2,045,400 random protein pairs, which did not include any detected interactions. We annotated proteins in the training sets with diverse features (for example, domains and localizations) and used a data-mining algorithm to identify sets of features that occurred together on numerous proteins. We assumed that a pair of feature sets, annotated to two proteins, could affect the chances of the proteins' interaction. We considered pairs of feature sets as one of five types of interaction evidence; the other types were gene coexpression, topology of the known PPI network, known orthologous interactions in model organisms and known paralogous interactions. Using a training set, we calculated probabilities of interaction from available evidence. Last, we combined these probabilities using a noisy-OR model[13]. With this model, a single type of evidence that strongly supports interaction can be sufficient for a high final probability. The final probability is not diminished if other evidence is absent, and it is increased if additional evidence is present (Online Methods and **Supplementary Note**).

We tested FpClass on gold-standard data sets in which positive cases comprised interactions detected in multiple experiments. We then assessed its performance on multiple PPI data sets detected by individual high-throughput or small-scale screens. PPIs in these data sets were enriched for low-degree proteins. FpClass outperformed previous PPI prediction methods on all such data sets. We validated 137 FpClass predictions using glutathione S-transferase (GST) pulldown, co-immunoprecipitation (Co-IP) or mammalian-membrane two-hybrid (MaMTH) assays[14]. Finally, we investigated the ability of FpClass to predict PPIs for orphans. We experimentally validated five predicted interactions involving these proteins. FpClass predicted 10,647 interactions for 1,089 such proteins at an estimated 60% FDR. We show that orphans have distinguishing properties that make detection and prediction of their interactions especially challenging.

RESULTS

Validation of PPI predictions by multiple approaches

First, we evaluated predictions using two gold standards: (i) a data set from Braun *et al.*[7] comprising 92 high-confidence PPIs (positive control) and 92 random protein pairs and (ii) our own data set. Braun *et al.*[7] evaluated five high-throughput assays for PPI detection: LUMIER, MAPPIT, a yeast two-hybrid (Y2H) assay, a protein-fragment complementation assay and wNAPPA. These methods identified between 21% and 36% of PPIs and up to 4% of non-interacting pairs[7]. Using the same data, FpClass predicted 90% of interactions and 1% of random pairs (**Fig. 2a**,**b** and **Supplementary Data 2**) at a probability threshold of 12%. Performance with other thresholds was similar: at 10% probability, predictions did not change, and at 25% probability, 86% of interactions and no random pairs were predicted by FpClass. These protein pairs were absent from the FpClass training data. Notably, the five methods together detected only 51% of interactions at 10% FDR[7]. Next, using fivefold cross-validation we evaluated FpClass on our gold-standard data set—our largest training set. This data set comprised 20,454 high-confidence PPIs detected by two independent experiments (positive control) and 2,045,400 random protein pairs (**Supplementary Data 3**). The random pairs excluded experimentally detected interactions but may have included undetected interactions. Using this data set, we evaluated different types of interaction evidence individually and in combination (**Fig. 2c**); the most useful types of evidence included topology, paralogs and orthologs. The combination of evidence identified more true interactions than individual evidence types did, especially at recalls greater than 0.5.

Next, we assessed performance on six PPI reference sets from multiple experimental assays including LUMIER, high-throughput mass spectrometry–based methods, small-scale screens and high-throughput Y2H assays (**Supplementary Table 1**). These reference sets were highly enriched ($P < 10^{-100}$) for low-degree

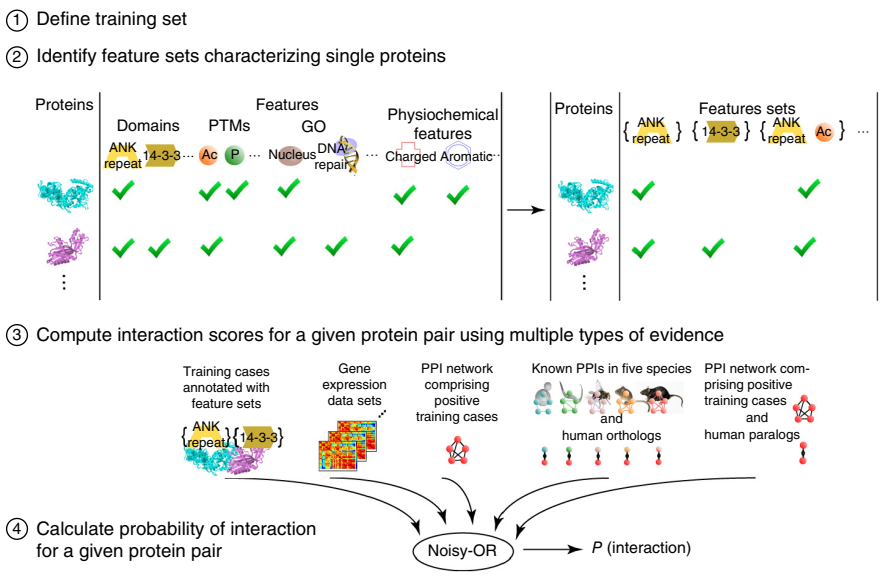

Figure 1 | FpClass workflow. The main steps for PPI prediction were assembling training data, identifying feature sets that may act cooperatively to affect interactions, calculating interaction scores and calculating interaction probability. (1) Our largest training set comprised 20,454 PPIs detected in at least two experiments (positive cases) and 2,045,400 random human protein pairs (negative cases). (2) Proteins were annotated with features including domains, post-translational modifications (PTMs), Gene Ontology (GO) terms and physicochemical properties. A data-mining algorithm was used to identify feature sets, comprising one or more of the above features, frequently occurring on proteins. Proteins were annotated with these feature sets. (3) Five types of evidence were used to compute interaction scores for a give protein pair (*i*,*j*). Feature-set pairs of (*i*,*j*) were analyzed for enrichment among positive training cases. Pearson correlations were calculated for genes encoding (*i*,*j*) using multiple gene-expression data sets. A network comprising positive training cases was used to assess the tendency of (*i*,*j*) to share interaction partners. Interactions of (*i*,*j*) orthologs in model organisms were identified, and sequence identity of (*i*,*j*) orthologs was assessed. Known interactions of (*i*,*j*) paralogs were similarly analyzed. (4) Probabilities of interaction for (*i*,*j*) were computed on the basis of interaction scores, integrated into a single probability using a noisy-OR model. This probability was calibrated using training data.

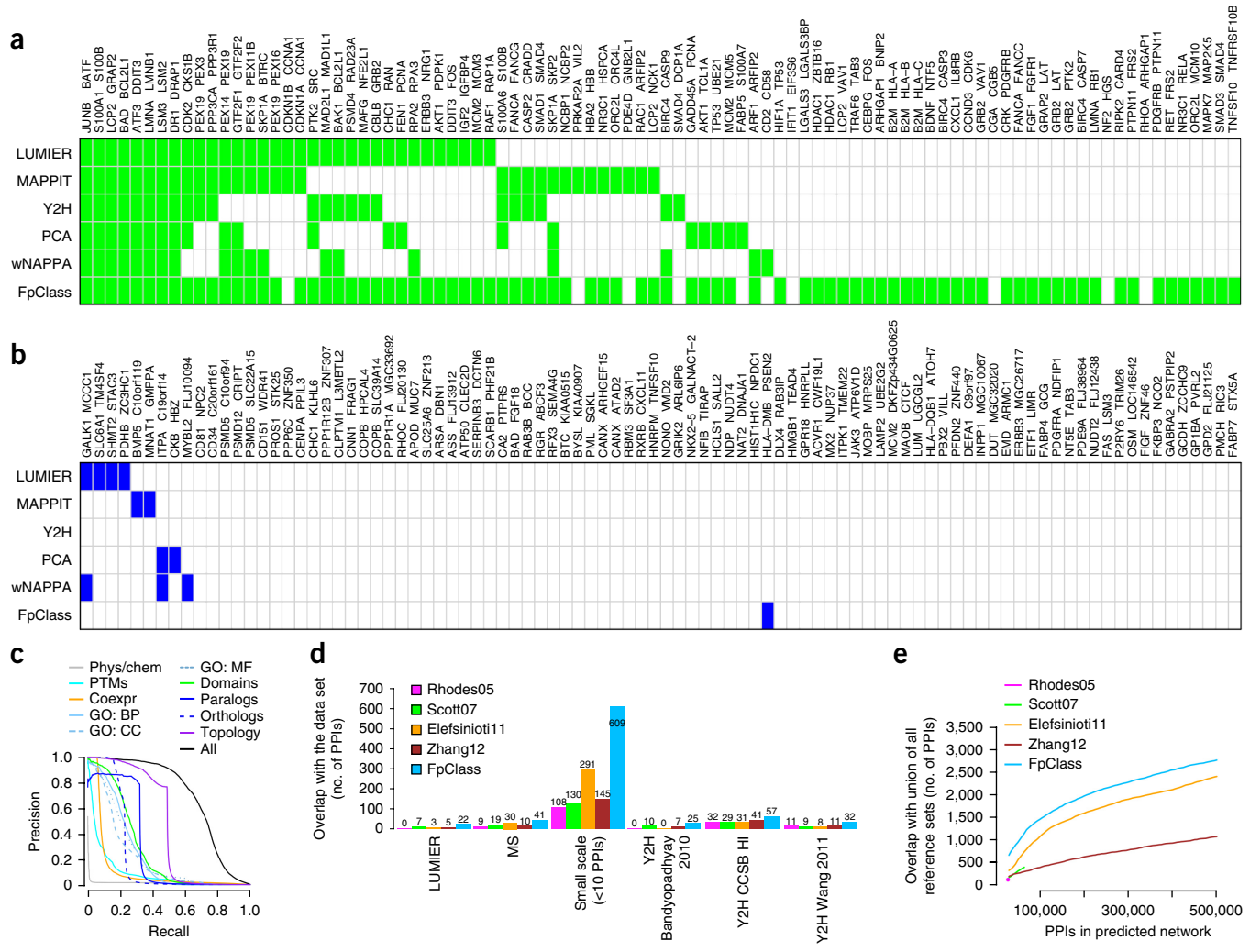

Figure 2 | Evaluation of FpClass using experimental PPI data sets. (**a**,**b**) PPI interactions detected or predicted by the five indicated experimental methods and FpClass, in the positive reference set (**a**) and random reference set (**b**) of Braun *et al.*[7]. Results of experimental methods are from Braun *et al.*[7]. (**c**) Precision (true positives/(true positives + false positives)) and recall (true positives/(true positives + false negatives)) of FpClass predictions, based on individual evidence types and all evidence, on a data set comprising 20,454 high-confidence PPIs and 2,045,400 random protein pairs. Individual evidence types included physicochemical protein properties (Phys/chem), gene coexpression (Coexpr), and Gene Ontology annotations of biological process (GO:BP), cellular compartment (GO:CC) and molecular function (GO:MF). (**d**–**f**) Interaction predictions of FpClass and the indicated prediction methods[11,12,15,16] compared against six experimental PPI reference sets (**Supplementary Table 1**). (**d**) Overlaps between the top 35,000 predictions of each computational method and each reference set. Numbers above the columns indicate the number of predicted PPIs that were present in a reference data set. (**e**) Overlaps between top predicted interaction networks of different sizes and the union of all reference sets. (**f**) Overlaps between top predicted interaction networks and reference sets from methods that report binary interactions (Y2H and LUMIER).

proteins (fewer than five known interactions). We also used these data sets to evaluate four other prediction methods[11,12,15,16]. We determined the overlaps between top predictions of each method, excluding any PPIs used in training, and these six reference sets (**Fig. 2d**). We then determined overlaps for increasing numbers of top predictions (**Fig. 2e**,**f**). FpClass had the highest overlaps with all references sets, and significantly lower FDRs than previous prediction methods ($P < 0.05$), estimated using the method of D'haeseleer and Church[17] (**Supplementary Fig. 1**).

Finally, we experimentally tested high- and medium-confidence predictions of FpClass (probabilities >0.25). We tested a total of 233 predictions: 5 by Co-IP, 30 by GST pulldown and 198 by MaMTH assays[14]. Each assay tested a different set of predictions. We confirmed 137 PPI predictions, including 64 interactions not reported in previous studies, corresponding to an FDR of 41% (**Fig. 3a**,**b**, **Supplementary Fig. 2**, **Supplementary Table 2** and **Supplementary Data 4**). FDRs from the three assays ranged from 39% to 60%. Differences in the FDRs may reflect differences in the test sets or assays or simply sampling effects. The most accurate estimate may come from the MaMTH assay because it tested the largest number of predictions.

The number of predictions we could experimentally test was small compared to the number of genome-wide predictions with probabilities >0.25. To estimate the FDR of genome-wide predictions, we used *in silico* testing (**Supplementary Fig. 1d**–**g**). This testing showed that a 60% FDR required predictions with

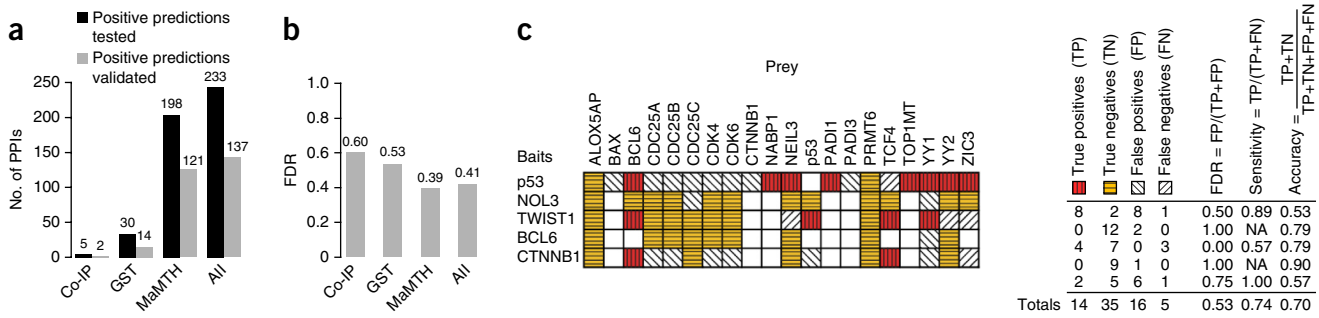

Figure 3 | Experimental validation of FpClass predictions. (**a**,**b**) 233 predictions with probabilities >0.25 were tested by Co-IP, GST pulldown or MaMTH assays. Validated predictions were detected in at least two experiments; each prediction was performed once. Numbers of interactions tested and confirmed by each method (**a**) and the resulting FDRs (**b**) are shown. (**c**) Results of testing 70 FpClass predictions with probabilities between 0.002 and 0.88 by GST pulldown assays. Probabilities >0.25 were assumed to indicate interaction. NA, not applicable.

probabilities greater than 0.47 rather than 0.25. At this probability threshold, FpClass predicted 250,498 interactions among 10,529 proteins. We considered these predictions to be high confidence and refer to them as the Fp60 network (**Supplementary Data 5**). Although the estimated FDR of 60% is fairly high, we consider this to be a reliable conservative estimate because it was confirmed in both *in vitro* and *in silico* testing.

For a more detailed assessment of performance, we also used GST pulldown assays to test 70 PPI predictions with a wide range of confidence levels (0.002–0.88) (**Fig. 3c**). Using a probability threshold of 0.25 for positive predictions, we obtained sensitivity of 74% and FDR of 53%. The rank-biserial correlation between predicted probabilities and detection status (detected or undetected) was 0.378 ($P = 0.0159$).

Predicting PPIs for proteins without known interactions

We applied FpClass to human proteins without known interactions (**Supplementary Data 6**). We defined these proteins as ones without detected PPIs from any experimental methods, including methods detecting protein complexes, according to the Interologous Interaction Database version 1.95 (**Supplementary Table 3**). We refer to these proteins as interactome orphans or d0 proteins (proteins with degree of 0) and to other proteins as dk proteins (proteins with degrees of $k > 0$).

Using GST pulldown assays, we tested FpClass predictions for five different baits, including p53, a well-characterized protein included in many previous PPI screens. Among the predicted preys, we randomly selected 20. These included some known interactors and, in particular, some interactome orphans. Five of six interactome orphans predicted to bind p53 were validated by GST pulldown assay: NABP1, NEIL3, TOP1MT, YY2 and PADI1. These five proteins share processes and diseases with p53, and also suggest new roles for p53. Both NABP1 and NEIL3 are nuclear proteins involved in DNA repair[18,19], a typical function of p53 (ref. 20). TOP1MT is a DNA topoisomerase involved in transcription and replication of mitochondrial DNA[21], and it is well established that p53 translocates to mitochondria in response to stress signals[22]. YY2 is a transcription factor that is thought to have arisen by retrotransposition of the *YY1* gene, and YY1 is a known interaction partner of p53 (ref. 23). PADI1 is a protein arginine deiminase that catalyzes the post-translational conversion of peptidylarginine to peptidylcitrulline. Emerging evidence supports a role for PADI1 in modulating the p53 response[24]. Top p53 predicted interactors included other interactome orphans and many known partners of p53 (**Supplementary Fig. 2d**).

We also investigated to what extent interactome orphans were included in proteome-wide, high-confidence predictions of FpClass. The Fp60 network included 1,089 interactome orphans, and their predicted interactions were supported by two types of independent, indirect evidence: (i) interactome orphans and their predicted partners often shared disease annotations ($P < 0.001$), and (ii) genes encoding interactome orphans and their partners were often regulated by the same drugs ($P < 0.001$).

Interactome orphans have unique properties

The most distinctive feature of interactome orphans is their recent origin (**Fig. 4a**). Interactome orphans include over 69% of mammal-specific and over 90% of primate-specific proteins ($P < 7 \times 10^{-31}$). In addition to their young age, interactome orphans have been recently evolving at almost twice the rate of dk proteins ($P < 5 \times 10^{-177}$) (**Fig. 4b**).

Gene Ontology (GO)[25] annotations suggest that interactome orphans are mainly involved in receptor- and metabolism-related functions (**Supplementary Fig. 3**). The most highly enriched GO terms were "signal transducer" and "receptor" ($P < 3 \times 10^{-5}$). Interactome orphans are not simply membrane proteins: surprisingly, they were not enriched for plasma-membrane localization (**Supplementary Fig. 3a**). Although localization is unknown for many interactome orphans, restricting analysis to annotated proteins still gave no plasma-membrane enrichment (**Supplementary Fig. 3d**) but gave enrichment for other terms: "extracellular region," "lipid metabolism" and "carbohydrate metabolism."

Top-enriched InterPro[26] annotations included receptors associated with smell and taste and domains related to transcription, extracellular-matrix structure and detoxification (**Supplementary Table 4**). Interactome orphans comprised about 50% of human receptors, one of five main drug-target classes[27] (**Supplementary Fig. 4a**,**b**). Orphans were enriched for the membrane structural class ($P < 5 \times 10^{-46}$), had short length (median of 332 residues versus 465 for other proteins; $P < 4 \times 10^{-233}$) and low structural disorder (median of 16% versus 22% for other proteins; $P < 2 \times 10^{-60}$) (**Supplementary Fig. 4c–e**). The length and disorder of interactome orphans were similar to those of extracellular proteins (345 residues median length, 15% median disorder).

Interactome orphan genes are expressed at lower levels and in fewer tissues than other genes. We analyzed expression of

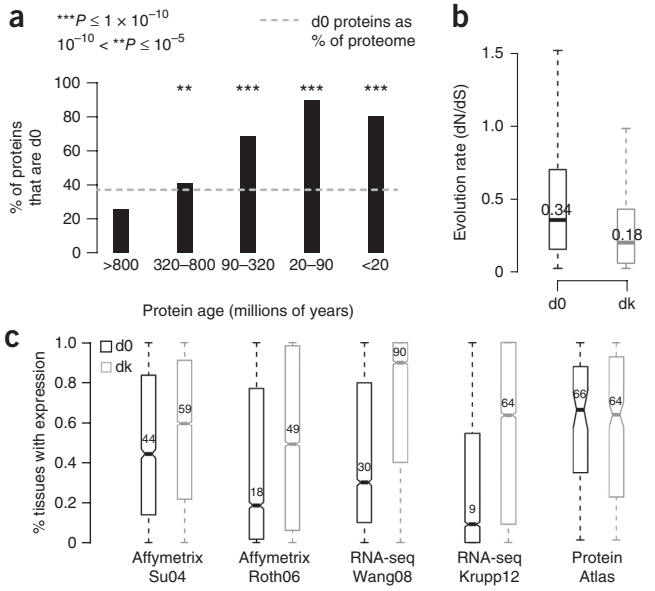

Figure 4 | Properties of d0 (orphan) proteins and genes. (**a**,**b**) Ages and evolution rates. (**a**) Prevalence of d0 proteins among proteins of various ages. *P* values were calculated by hypergeometric tests and adjusted for multiple testing using FDR. The dashed line indicates the percentage of human proteins that are d0. (**b**) Evolution rates of d0 and dk proteins; dN and dS are the numbers of nonsynonymous and synonymous substitutions per site, respectively. Values of dN and dS were determined from a comparison of human and chimpanzee orthologs. (**c**) Tissue specificity: percentages of tissues where d0 and dk genes (proteins) are expressed. Data were taken from Su *et al.*[28] (gene expression microarrays, 79 tissues), Roth *et al.*[29] (gene expression microarrays, 65 tissues), Wang *et al.*[30] (RNA-seq, 10 tissues), Krupp *et al.*[31] (RNA-seq, 11 tissues), and Uhlen *et al.*[32] (Human Protein Atlas, 83 tissues). For comparison of d0 and dk: $P < 0.0001$ using Affymetrix Su04, Affymetrix Roth06, RNA-seq Wang08 and RNA-seq Krupp12; $P = 0.751$ using Protein Atlas. *P* values were calculated by two-sided Mann-Whitney *U*-tests, as the data were not normally distributed. Box plots show minimum and maximum values (whiskers) and the first quartile, median and third quartile (hinges).

interactome orphan genes and proteins in healthy human tissues using data from gene expression microarrays (Affymetrix)[28,29], RNA-seq[30,31] and the Human Protein Atlas[32]. In gene expression data sets, interactome orphans had significantly lower median and maximum expression ($P < 2 \times 10^{-50}$) than other genes (**Supplementary Fig. 5**) and were highly tissue specific ($P < 4 \times 10^{-26}$) (**Fig. 4c**).

Features of interactome orphans may at least partially explain their paucity of detected interactions, including for the following reasons. First, interactome orphans are tissue specific and may be absent from cell types used in PPI screens. About 70% of interactome orphan proteins have not been detected by experimental assays—evidence for their existence comes from detection of mRNA transcripts or genomic analysis[33]. Second, interactome orphans are frequently involved in signaling; thus their interactions may occur only under specific conditions. However, PTMs are deficient among interactome orphans (**Supplementary Table 5**). Third, common detection methods such as Y2H and tandem affinity purification may be hindered by features of interactome orphans such as extracellular localization and low expression. Many orphans are extracellular and may be quickly exported from cells. We investigated this by analyzing results of two large human Y2H screens (**Supplementary Table 6**). In each screen, we identified proteins that were tested but for which no interactions were detected; these proteins were significantly enriched for features of interactome orphans. Also, we considered whether methods that report binary interactions (for example, Y2H) tend to miss different proteins from those missed by methods that detect protein complexes (for example, tandem affinity purification); we found that both assays tend to miss proteins similar to interactome orphans (**Supplementary Table 7**). Fourth, interactome orphans may have been excluded from assays owing to research bias. About 27% of interactome orphans had no orthologs in model organisms, compared to 3% of dk proteins (**Supplementary Table 8**). This reduces the chances that interactome orphans have known functions or disease associations. Researchers are less likely to study proteins without such information[6].

Features of interactome orphans make interaction detection and prediction more challenging but not impossible. Low-degree proteins ($k < 5$) in the Interologous Interaction Database (I2D)[34] have features similar to those of interactome orphans (**Supplementary Table 9**), whereas higher degree proteins ($k \geq 5$) do not. This suggests that features of interactome orphans reduce the chances of detecting interactions.

DISCUSSION

Two main concerns are commonly raised about the known human interactome: the high false positive rate of reported interactions and, to a lesser extent, the large number of undiscovered interactions. We show that proteins missing from the known interactome are not a random subset of the proteome. This bias poses an important challenge for interactome research as it suggests that the current understanding of the interactome (for example, its size, degree distribution and mathematical model) may be unreliable and that disease mechanisms involving recently evolved proteins remain poorly understood.

The focus of PPI prediction methods has often been on prediction accuracy: although an important factor, there may be limited value in achieving high accuracy mostly for proteins that have been extensively studied while certain categories of proteins are largely missing from the known interactome. PPI prediction methods have great potential for addressing interactome biases, but their development and evaluation may need to be modified—for interactome orphans, retrospective validations are not possible and reagents are frequently unavailable[6]; thus, validating orphan interactions requires greater effort than validating other interactions.

Predicting and detecting PPIs of interactome orphans remains an unsolved problem. FpClass predicted high-confidence PPIs for 1,089 interactome orphans, leaving 4,500 orphans without PPIs. Interactome orphans with and without high-confidence predictions are largely similar (**Supplementary Table 10**). Improved prediction and detection of orphan interactions faces a number of challenges including sparse annotation of proteins, research bias and technical difficulties of studying recent, tissue-specific proteins. Interactome orphans may be ideal drug targets because many of them are receptors and are highly tissue specific. Displacement of PPIs is a growing strategy for targeting disease-associated proteins (for example, displacement of the interaction between p53 and MDM2 or between p53 and Twist1; refs. 35,36),

and the ability to identify orphan interactions may help address the pharmacological targeting of these proteins.

We have introduced strategies for addressing the gap in interactome mapping, and we have applied FpClass to *de novo* proteome-wide PPI prediction. Although we achieved low FDR and higher overlap with existing PPI data sets than have other computational methods, the primary contribution of FpClass may be in reformulating and partially addressing the priorities for PPI prediction. The high validation rate for MaMTH[14] suggests that FpClass could help guide high-throughput screening in a combined computational-experimental approach to interactome mapping[10]. FpClass code is available as **Supplementary Software** and from our website (http://www.cs.utoronto.ca/~juris/data/fpclass/), which will host updates.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research was supported by the grants from Genome Canada via the Ontario Genomics Institute, Ontario Research Fund (GL2-01-030, RE-03-020 to I.J.), the Canadian Institutes for Health Research (#99745, #93579 to I.J., A.J.), the Natural Sciences Research Council (#203475 to I.J.), US Army Department of Defense W81XWH-12-1-0501 (to I.J.), the Italian Association for Cancer Research, the Friuli Venezia-Giulia and CRO 5xmille Intramural Grant (to R.M.), the Friuli Venezia-Giulia Exchange Program (to C.P.), the Ontario Genomics Institute (#303547 to I.S.), the Canadian Institutes of Health Research (Catalyst-NHG99091, PPP-125785 to I.S.), the Canadian Cystic Fibrosis Foundation (#300348 to I.S.), the Canadian Cancer Society (2010-700406 to I.S.), Genentech and University Health Network (GL2-01-018 to I.S.), US National Institutes of Health (NIH) PO1/PPG grant 01CA0099031 (to G.B.M., I.J.) and NCI R21 CA126700 (to Z.D., G.B.M.). Computational resources were supported by grants from the Canada Foundation for Innovation (CFI #12301, #203373, #29272, #22540a, #30865) and IBM (to I.J.). I.J. is supported by the Canada Research Chair program. This research was also supported by the University of Toronto McLaughlin Centre and the Ontario Ministry of Health and Long-Term Care (OMOHLTC). The views expressed do not necessarily reflect those of the OMOHLTC. We thank M. Vidal, D. Hill, F. Roth and the Center for Cancer Systems Biology (Dana-Farber Cancer Institute) for prepublication release of protein interaction data, funded by NIH NHGRI grant R01 HG001715.

AUTHOR CONTRIBUTIONS

M.K. and I.J. conceived of the project. M.K. developed the algorithm and executed computational analyses and validation. Additional validation and assay-related analyses were executed by C.P., C.C., Y.N., F.V. and F.B.-C. R.M., I.S., A.J. and G.B.M. provided guidance for biological validation experiments that were executed by F.P., A.L.S., H.L., C.P., T.N. and Z.D. M.K. and I.J. wrote the initial manuscript, and all authors were involved in results presentation, discussion and preparation of the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Cusick, M.E. *et al.* Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39–46 (2009).
2. Pastrello, C. *et al.* Integration, visualization and analysis of human interactome. *Biochem. Biophys. Res. Commun.* **445**, 757–773 (2014).
3. Bork, P. *et al.* Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292–299 (2004).
4. Stumpf, M.P. *et al.* Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**, 6959–6964 (2008).
5. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
6. Edwards, A.M. *et al.* Too many roads not taken. *Nature* **470**, 163–165 (2011).
7. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2009).
8. Brückner, A., Polge, C., Lentze, N., Auerbach, D. & Schlattner, U. Yeast two-hybrid, a powerful tool for systems biology. *Int. J. Mol. Sci.* **10**, 2763–2788 (2009).
9. Wodak, S.J., Pu, S., Vlasblom, J. & Séraphin, B. Challenges and rewards of interaction proteomics. *Mol. Cell. Proteomics* **8**, 3–18 (2009).
10. Schwartz, A.S., Yu, J., Gardenour, K.R., Finley, R.L. Jr. & Ideker, T. Cost-effective strategies for completing the interactome. *Nat. Methods* **6**, 55–61 (2009).
11. Rhodes, D.R. *et al.* Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* **23**, 951–959 (2005).
12. Scott, M.S. & Barton, G.J. Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics* **8**, 239 (2007).
13. Kim, J.H. & Pearl,, J. in *Proc. IJCAI* 190–193 (Morgan Kaufmann, 1983).
14. Petschnigg, J. *et al.* The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells. *Nat. Methods* **11**, 585–592 (2014).
15. Elefsinioti, A. *et al.* Large-scale *de novo* prediction of physical protein-protein association. *Mol. Cell. Proteomics* **10**, M111.010629 (2011).
16. Zhang, Q.C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).
17. D'haeseleer, P. & Church, G.M. in *Proc. IEEE Comput. Syst. Bioinform. Conf.* 216–223 (IEEE, 2004).
18. Kang, H.S. *et al.* NABP1, a novel RORg-regulated gene encoding a single-stranded nucleic-acid-binding protein. *Biochem. J.* **397**, 89–99 (2006).
19. Krokeide, S.Z. *et al.* Human NEIL3 is mainly a monofunctional DNA glycosylase removing spiroimindiohydantoin and guanidinohydantoin. *DNA Repair (Amst.)* **12**, 1159–1164 (2013).
20. Menendez, D., Inga, A. & Resnick, M.A. The expanding universe of p53 targets. *Nat. Rev. Cancer* **9**, 724–737 (2009).
21. Wang, W. *et al.* Identification of rare DNA variants in mitochondrial disorders with improved array-based sequencing. *Nucleic Acids Res.* **39**, 44–58 (2011).
22. Vaseva, A.V. & Moll, U.M. The mitochondrial p53 pathway. *Biochim. Biophys. Acta* **1787**, 414–420 (2009).
23. Gordon, S., Akopyan, G., Garban, H. & Bonavida, B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* **25**, 1125–1142 (2006).
24. Tanikawa, C. *et al.* Regulation of protein citrullination through p53/PADI4 network in DNA damage response. *Cancer Res.* **69**, 8761–8769 (2009).
25. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
26. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
27. Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.* **5**, 821–834 (2006).
28. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
29. Roth, R.B. *et al.* Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **7**, 67–80 (2006).
30. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
31. Krupp, M. *et al.* RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics* **28**, 1184–1185 (2012).
32. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).
33. The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142–D148 (2010).
34. Brown, K.R. & Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* **8**, R95 (2007).
35. Piccinin, S. *et al.* A "twist box" code of p53 inactivation: twist box: p53 interaction promotes p53 degradation. *Cancer Cell* **22**, 404–415 (2012).
36. Hupp, T.R., Hayward, R.L. & Vojtesek, B. Strategies for p53 reactivation in human sarcoma. *Cancer Cell* **22**, 283–285 (2012).

ONLINE METHODS

PPI predictions. Our approach for predicting PPIs comprised four stages: defining a training set, identifying feature sets of single proteins, computing interaction scores of protein pairs, and computing probabilities of interaction on the basis of these scores (**Fig. 1**).

First, we assembled a set of protein pairs for training the classifier. The largest data set contained ~2.1 million protein pairs: 20,454 PPIs detected in at least two independent experiments and 2,045,400 random protein pairs, excluding any experimentally identified interactions. After training the classifier we calculated interaction probabilities for all human protein pairs.

In the second stage we identified sets of protein features (for example, {domain = TSP type-1, PTM = C-linked glycosylation}) that may work together to affect a protein's interactions. We assumed that some of these sets correspond to frequently co-occurring protein features. To find these sets, we first annotated human proteins with the following feature types: domains, GO terms, PTMs, and physicochemical features (for example, charge, aromaticity). Next, using a frequent pattern-mining algorithm, we identified sets of protein features that co-occurred more frequently than expected by chance. Resulting feature sets contained one or more features, and the features could be of the same or different types (for example, {GO component = nucleus, GO component = cytoplasm}, {domain = SH3, PTM = phosphorylation}).

In the third stage, we computed interaction scores for pairs of human proteins. Five types of scores were computed for a protein pair; each score indicated the level of interaction support from one type of evidence.

1. Pairs of feature sets from individual proteins. We used pairs of feature sets to help predict interactions, as previous studies used pairs of protein domains[37]. First, for all pairs of feature sets (from step 2), we calculated log-odds ratios, indicating their over- or underrepresentation among positive training cases. Positive values indicated overrepresentation, whereas negative values indicated underrepresentation. Then, for a given pair of proteins (i,j), we considered all pairings between feature sets of i and j and assigned to the pair the log-odds ratio with the highest magnitude.

2. Gene expression data. If the genes encoding proteins i and j were present in a gene expression data set, we calculated the Pearson correlation of their expression profiles.

3. Topology data. If proteins (i,j) were both present in the training data, three scores were calculated to indicate the tendency of i and j to share interaction partners.

4. Orthology data. If proteins (i,j) had interacting orthologs (i',j') in a model organism, sequence identities were determined between i and i' and between j and j'. The lower of these two values was used as a score for proteins (i,j). Such scores were determined from five model organisms: yeast, worm, fly, mouse and rat.

5. Paralogy data. If proteins (i,j) had interacting paralogs (i',j') in the training data, the lower of the two sequence identities *identity*(i,i') and *identity*(j,j') was used as a score.

In the fourth stage, interaction scores were used to calculate a single probability of interaction. First, each score $s_{(i,j),k}$ for protein pair (i,j) was converted to a probability of interaction $P(I_{(i,j)} \mid s_{(i,j),k})$ by calculating the fraction of positive training cases among all training cases with scores $\geq s_{(i,j),k}$. Resulting probabilities were combined into a single probability using a noisy-OR model[13]: $p_{(i,j)} = 1 - \Pi^{n}_{k=1}(1 - P(I_{(i,j)} \mid s_{(i,j),k}))$, where n is the number of scores for pair (i,j). A final probability of interaction, $P(I_{(i,j)} \mid p_{(i,j)})$, was calculated in two steps. First, a probability was calculated as the fraction of positive training cases among all training cases exceeding $p_{(i,j)}$. This probability was then adjusted to account for the fact that the frequency of positive cases in training data, 1:100, was higher than the frequency of interactions among human protein pairs, which we assumed to be 1:600. Additional details are available in the **Supplementary Note**. The prediction algorithm is available at http://www.cs.utoronto.ca/~juris/data/fpclass/.

Shared features of interactome orphans and predicted partners. Using the Fp60 network, we tested whether interactome orphans and their predicted partners frequently shared two types of features: (i) disease annotations from literature[38–40] and (ii) drugs significantly regulating the proteins' genes[41]. For each feature, we determined the number, N_O, of interactome orphans sharing an annotation with at least one of their partners. We calculated a P value for N_O on the basis of 1,000 randomizations; in each randomization, annotations of interactome orphans (in the Fp60 network) were randomly permutated (i.e., each interactome orphan was assigned annotations of another interactome orphan), and the number N_R of interactome orphans sharing an annotation with a partner was determined. A P value was the fraction of randomizations with $N_R > N_O$.

Independence of training and test sets. We tested FpClass on our gold-standard data set (~2.1 million protein pairs) by five-fold cross-validation. To test FpClass on gold-standard data from Braun *et al.*[7], we trained on different data sets, depending on whether the protein pair being predicted was in our gold-standard data. If the protein pair was in our data set, we trained on a subset of our data set that excluded the protein pair. This subset was used in one of the cross-validation folds as a training set. If the protein pair was not in our gold-standard data, we trained on our entire data set. We used a similar approach for testing FpClass on other data.

GST pulldown experiments. GST pulldown experiments were done as previously described[35]. Briefly, indicated clones were *in vitro* translated (IVT) with [^{35}S]methionine (PerkinElmer) using the TNT System (Promega). IVT proteins were then incubated together with either GST-human p53 or GST-only for 2 h at 4 °C in binding buffer (20 mM Tris-Cl, pH 8.0, 150 mM NaCl, 1 mM EDTA, 0.1% Nonidet-P40, 1 mM PMSF and Complete (Roche) protease inhibitors) plus glutathione Sepharose resin (GE Healthcare). After extensive washings, bound proteins were separated by SDS-PAGE. Gels were stained with Coomassie brilliant blue, dried and exposed to a Cyclone Phosphor Imager (Packard) first and then to films (Kodak). Each experiment was repeated at least twice.

Co-immunoprecipitation experiments. HEK cells were obtained from the laboratory of H. McNeill (Lunenfeld-Tanenbaum Research Institute) and were tested for mycoplasma infection using MycoAlert (Lonza). HEK293 cells were maintained in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% FBS and antibiotics. Transfection was performed using Metafectene Pro (Biontex Laboratories) according to the

manufacturer's instructions using pcDNA3-Myc or HA-tagged Asc or HA-Ripk1 vectors. Cells were harvested in 0.1% SDS RIPA lysis buffer. Whole lysates were incubated overnight with 2 µg anti-HA (sc-7392; Santa Cruz Biotechnology), 2 µg anti-ASC (sc-30153; Santa Cruz Biotechnology) or 4 µg anti-Myc (45A-Z; Immunology Consultants Laboratory) antibody. Lysates were subsequently incubated with protein Sepharose beads (True Blot, Rockland), washed and centrifuged, and the protein-antibody complexes that were recovered were subjected to western blot analysis. Blots were hybridized with 1:400 anti–caspase 8 (9746; Cell Signaling), 1:500 anti-PAK1 (NBP1-85802; Novus Biologicals) and 1:200 anti-HA (sc-7392; Santa Cruz Biotechnology) antibodies followed by an appropriate HRP-conjugated secondary antibody (1:1,000; anti-rabbit or anti-mouse HRP True Blot, Rocklad, or 1:5,000 anti-mouse HRP; Bio-Rad) using Western Lightning Plus-ECL (PerkinElmer). Each experiment was repeated three times.

Defining the human proteome and interactome orphans. We defined the human proteome as comprising 19,698 proteins from the UniProt database[33] release-2011_05, with unique Entrez[42] IDs and evidence status other than "Uncertain." We defined interactome orphans as human proteins without experimentally detected interactions in the Interologous Interaction Database (I2D)[34] version 1.95. I2D integrates known interactions from existing curated databases (for example, BIND, BioGRID, DIP, HPRD, IntAct, MINT) and high-throughput experimental data. Although I2D includes PPIs predicted from orthologous interactions in other organisms, we did not include such PPIs in our analysis. We considered only whether human proteins were involved in experimentally detected interactions (i.e., were part of the known human interactome) or were not involved in any such interactions (i.e., interactome orphans).

Identifying features of interactome orphans. *Protein age.* We estimated protein age as the oldest divergence time between humans and species with orthologs of the protein. We obtained orthologs of human proteins in 55 species using Ensembl BioMart (Ensembl Genes 65). We divided these species into five groups: apes including gibbons, primates other than apes, mammals other than primates, vertebrates other than mammals, and eukaryotes other than vertebrates. Using the TimeTree database[43] (12 July 2011) we defined approximate maximum divergence times between *Homo sapiens* and these five groups of species: <20 million years for apes, 20–90 million years for primates other than apes, 90–320 million years for mammals other than primates, 320–800 million years for vertebrates other than mammals, and >800 million years for eukaryotes other than vertebrates. We estimated the age of a protein by identifying groups of species with orthologs of the protein and taking the maximum divergence time between these groups and *H. sapiens*. The idea of our approach is based on Toll-Riera *et al.*[44].

Evolution rate. We downloaded evolution rates, dN/dS (nonsynonymous vs. synonymous substitutions), of human proteins relative to chimpanzee using Ensembl BioMart (Ensembl Genes 65).

Gene and protein expression in tissues. We obtained gene expression data for healthy human tissues from two microarray (Affymetrix)[28,29] and two RNA-seq[30,31] studies. We processed microarray data sets by normalizing with the MAS5 algorithm, averaging experimental replicates, and if a gene had multiple probe sets, we kept the one with the highest variance. We considered genes to be expressed in a tissue if their levels were >200 (ref. 28). We used RNA-seq data provided by authors[28,29] and considered genes to be expressed in a tissue if their reads per kilobase of transcript per million mapped reads (RPMKs) were at least 1 (ref. 45). We used protein expression data from the Human Protein Atlas (version 11.0)[32] and considered proteins to be expressed in a tissue if their levels were not "Negative" or "None."

Gene Ontology (GO). We annotated genes with a subset of terms from the goslim_generic set (http://www.ebi.ac.uk/QuickGO). We downloaded associations of human genes with goslim_generic categories on 6 June 2012.

Protein structure. We obtained domains of human proteins from InterPro[26] release 30.0 and predictions of SCOP structural classes from the AutoPSI database[46] (9 June 2012). We obtained protein lengths from UniProt release-2012_07. We predicted protein disorder using DISOPRED version 2.3 (ref. 47).

Analysis of interactome orphan features. We calculated enrichment of a given feature (for example, GO terms) among interactome orphans using cumulative hypergeometric probability and adjusted for multiple testing using FDR. We calculated the significance of differences between d0 and dk proteins using the Mann-Whitney *U*-test.

37. Sprinzak, E. & Margalit, H. Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.* **311**, 681–692 (2001).
38. Zhang, Y. *et al.* Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Med. Genomics* **3**, 1 (2010).
39. Osborne, J.D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10** (suppl. 1), S6 (2009).
40. Davis, A.P. *et al.* The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.* **39**, D1067–D1072 (2011).
41. Kotlyar, M., Fortney, K. & Jurisica, I. Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods* **57**, 499–507 (2012).
42. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **35**, D26–D31 (2007).
43. Hedges, S.B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).
44. Toll-Riera, M. *et al.* Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).
45. Barshir, R. *et al.* The TissueNet database of human tissue protein-protein interactions. *Nucleic Acids Res.* **41**, D841–D844 (2013).
46. Birzele, F., Gewehr, J.E. & Zimmer, R. AutoPSI: a database for automatic structural classification of protein sequences and structures. *Nucleic Acids Res.* **36**, D398–D401 (2008).
47. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. & Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).

doi:10.1038/nmeth.3178

Erratum: *In silico* prediction of physical protein interactions and characterization of interactome orphans

Max Kotlyar, Chiara Pastrello, Flavia Pivetta, Alessandra Lo Sardo, Christian Cumbaa, Han Li, Taline Naranian, Yun Niu, Zhiyong Ding, Fatemeh Vafaee, Fiona Broackes-Carter, Julia Petschnigg, Gordon B Mills, Andrea Jurisicova, Igot Stagljar, Gordon B Mills, Roberta Maestro & Igor Jurisica

***Nat. Methods*; doi:10.1038/nmeth.3178; corrected online 10 December 2014.**

In the version of this article initially published online, an author (G.B.M.) was incorrectly listed twice. The error has been corrected for the print, PDF and HTML versions of this article.