D536–D541 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 29 October 2015*
doi: 10.1093/nar/gkv1115

Integrated interactions database: tissue-specific view of the human and model organism interactomes

Max Kotlyar[1], Chiara Pastrello[1], Nicholas Sheahan[2] and Igor Jurisica[1,3,*]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, M5G 1L7, Canada, [2]School of Computing, Queen's University, Kingston, ON, K7L 2N8, Canada and [3]Departments of Medical Biophysics and Computer Science, University of Toronto, Toronto, ON, M5S 1A4, Canada

Received September 14, 2015; Accepted October 13, 2015

ABSTRACT

IID (Integrated Interactions Database) is the first database providing tissue-specific protein–protein interactions (PPIs) for model organisms and human. IID covers six species (*S. cerevisiae* (yeast), *C. elegans* (worm), *D. melonogaster* (fly), *R. norvegicus* (rat), *M. musculus* (mouse) and *H. sapiens* (human)) and up to 30 tissues per species. Users query IID by providing a set of proteins or PPIs from any of these organisms, and specifying species and tissues where IID should search for interactions. If query proteins are not from the selected species, IID enables searches across species and tissues automatically by using their orthologs; for example, retrieving interactions in a given tissue, conserved in human and mouse. Interaction data in IID comprises three types of PPI networks: experimentally detected PPIs from major databases, orthologous PPIs and high-confidence computationally predicted PPIs. Interactions are assigned to tissues where their proteins pairs or encoding genes are expressed. IID is a major replacement of the I2D interaction database, with larger PPI networks (a total of 1,566,043 PPIs among 68,831 proteins), tissue annotations for interactions, and new query, analysis and data visualization capabilities. IID is available at http://ophid.utoronto.ca/iid.

INTRODUCTION

Cellular processes are carried out through protein–protein interactions (PPIs); identifying these interaction networks enables a better understanding of the mechanisms behind different phenotypes. Known PPI networks have proven valuable for many applications, including prediction of gene function (1,2), identification of disease genes (3,4) and drug discovery (5,6).

However, the usefulness of known networks is limited by several factors: most interactions lack context information (e.g. location and time), many interactions are missing (high false negative rate) and many are false positives. These limitations are especially acute for model organism interactomes. This is a key problem since the tasks where networks may be most beneficial, such as drug discovery, are primarily studied in these organisms. Several types of context information including tissue, subcellular localization and disease associations are available for some human PPIs from the HIPPIE database (7). Tissues for human PPIs are also available from the TissueNet database (8) and several other studies (9–11), though the reliability of tissue assignments is unclear. The ComPPI database (12) provides subcellular localizations for human and model organism PPIs. Missing interactions are an important problem for human and model organism interactomes. The human interactome, estimated at up to 650,000 PPIs (13), may be less than one-third complete. Databases of experimentally detected, curated human PPIs (14–19) report up to approximately 150,000 interactions. Online resources, such as iRefWeb (20), STRING (21) and ConsensusPathDB (22), integrate these databases to obtain about 240,000 human PPIs. This number can be further extended with predicted PPIs (23–25) but databases tend to focus on either detected or predicted interactions, though STRING (21) includes predictions of functional interactions. Interactions of non-human species are available in many PPI databases, but these interactomes, with the exception of yeast, are likely far less complete than human. The largest number of detected PPIs available for a non-yeast model organism is about 30,000 for mouse. The problem of false positives may be easier to assess and address than the number of missing interactions. Several PPI databases have developed confidence scores for interactions, and benchmarked their scores against gold standard data sets. However, gold standard data sets, typically comprising interactions detected by multiple small-scale screens, may have biases (25), and can be difficult to generate for organisms with few well-studied interactions.

IID aims to reduce the limitations of human and especially model organism PPI networks, making these networks more useful for experimental studies. Typically, animal models are used to investigate the roles of specific genes

*To whom correspondence should be addressed. Tel: +1 416 581 7437; Email: juris@ai.utoronto.ca

or proteins in disease, with the assumption that the roles may be similar in humans. Comparing human and model organism networks can indicate if this is the case; proteins under investigation may play similar roles in human disease if their interactions are largely conserved, and occur in the same tissues, in the two species. IID provides this information for human and five model organism networks (*Saccharomyces cere*visiae (yeast), *Caenorhabditis elegans* (worm), *Drosophila melonogaster* (fly), *Rattus norvegicus* (rat), *Mus musculus* (mouse)), and annotates interactions in each species except yeast with up to 30 tissues. Typical input to IID comprises a set of protein or gene IDs, and one or more tissues and species. If input proteins are not from the selected species, IID automatically includes their orthologs in the query, unless specified otherwise by the user. It returns interactions of the input proteins and their orthologs occurring in any of the specified tissues and species. Alternatively, users can specify that interactions should be conserved across tissues, species or both. To reduce the number of missing interactions, especially for model organisms, IID includes orthologous interactions, generated by mapping experimentally detected PPIs in any of the six species (human and five model organisms) to orthologous protein pairs in the remaining five species. IID also includes high-confidence predicted PPIs from genome-wide prediction studies (23–26). This reduces the number of missing interactions and can serve a similar role as confidence scores for detected interactions that have also been predicted by one or more studies, and thus are likely more reliable. Users can exclude interactions based on evidence type: experimental detection, orthology or prediction.

MATERIALS AND METHODS

Data sources

PPIs. Experimentally detected PPIs were downloaded from seven databases: BioGRID (14) 3.4.125, DIP (19) 2015-01-01, HPRD (17) Release 9, I2D (27) 2.3, InnateDB (18) 2015-05-23, IntAct (15) 2015-06-13 and MINT (16) 2013-03-26. Four sets of predicted PPIs were obtained: predictions from Rhodes *et al.* (26) with a likelihood ratio cut-off of 381, predictions from Elefsinioti *et al.* (23) with probabilities greater than 0.7, predictions from Zhang *et al.,* (24) with likelihood ratio cut-off of 600 and predictions Kotlyar *et al.* (25) with a false discovery rate less than 0.6.

Gene expression. Eight gene expression data sets were downloaded from NCBI GEO (28): GSE10246, GSE1133, GSE23328, GSE24207, GSE3526, GSE7307, GSE7763 and GSE9485. All data sets were normalized using the mas5 function in the *affy* package (29) in R. In each data set, disease tissues were removed, replicates were averaged and probeset IDs were mapped to Entrez Gene IDs. If a gene was represented by multiple probesets, the one with the highest variance was chosen.

Protein expression. Protein expression data sets were downloaded from Human Protein Atlas (30) version 13 and PaxDb (31) version 4.

Orthologs. Orthologs were downloaded from HomoloGene (32) build 68.

Mapping between gene and protein IDs

Mappings between various gene and protein IDs were based on UniProt (33) release 2015_06.

Assigning interactions to tissues

An interaction was assigned to a tissue if its two proteins or encoding genes were expressed in the tissue. A gene was considered expressed in a tissue if its mas5 normalized expression was above 200, as in Bossi *et al.,* (9). A protein was considered expressed in a tissue if its level based on Human Protein Atlas (30) was anything other than 'Not detected' or its level based on PaxDb (31) was greater than 0.

Generating orthologous interactions

Orthologous interactions were generated by mapping experimentally detected PPIs in each of the six IID species, to pairs of Homologene (32) orthologs in the other five species, if such orthologs were available.

Counting graphlets

Graphlet counts were calculated using Orca (34).

RESULTS

IID contents

IID has a total of 1,566,043 PPIs and 68,831 proteins for six species (*S. cerevisiae* (yeast), *C. elegans* (worm), *D. melonogaster* (fly), *R. norvegicus* (rat), *M. musculus* (mouse) and *H. sapiens* (human))—corresponding to a 74% increase in PPIs and a 10% increase in proteins over I2D version 2.3 (Table 1). Interactions are based on three types of evidence: experimental detection, orthology and *in silico* prediction. Predictions are primarily available for human, and represent 78% of the human network. Orthologous interactions are most important for model organisms other than yeast, representing between 43% and 97% of interactions in these networks.

For five species other than yeast, IID annotates interactions with up to 30 tissues. Available tissues for each species are shown in Supplementary Table S1. Most PPIs (46–92%) are annotated with at least one tissue (Figure 1). These PPIs are rarely tissue-specific; about two-thirds are annotated to more than half of the tissues in a species (Figure 2). Surprisingly, the trends are very similar across species. Similarly, all tissues in a species are associated with over 55% of PPIs in the species (Figure 3).

Human PPI networks in all 29 tissues are well conserved in mouse, and to a lesser extent in other model organisms (Figure 4). Over 85% of human experimentally detected PPIs in a tissue can be mapped to orthologous protein pairs in mouse, and over half of these orthologous pairs are annotated to the same tissue in mouse.

Table 1. Numbers of interactions in IID compared with I2D 2.3

Database	IID				I2D		
Species/Prediction	Experimental	Orthologous	Predicted	Total	Experimental	Orthologous	Total
Human	204,474	57,829	664,643	850,636	183,524	55,985	228,847
Mouse	29,273	204,305	-	225,247	19,090	190,049	203,114
Rat	5,665	168,137	-	173,802	4,178	116,649	119,527
Fly	59,200	43,037	-	100,316	53,325	45,849	97,967
Worm	13,678	28,567	-	41,544	11,555	39,606	50,486
Yeast	144,526	6,996	61,720	176,351	191,673	12,810	200,587

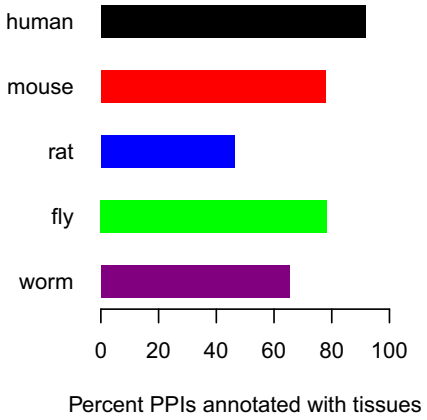

Figure 1. Percentages of PPIs annotated with tissues.

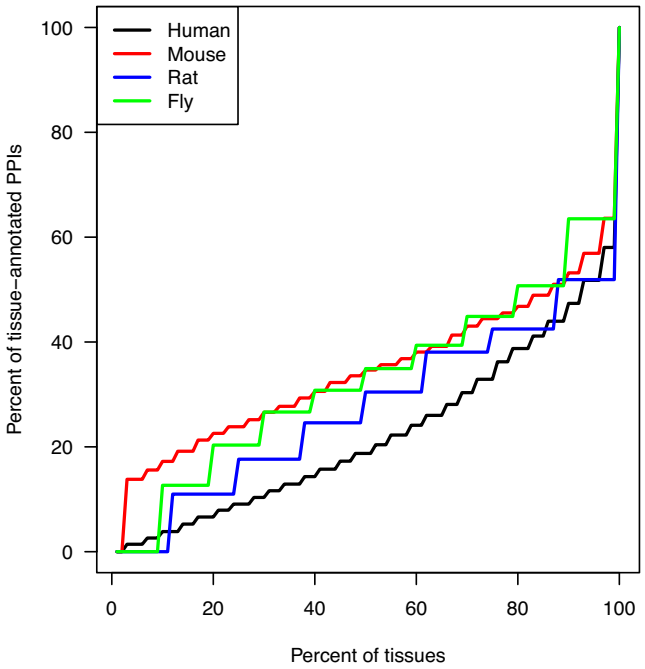

Figure 2. Tissue specificity of PPIs (i.e. are most PPIs annotated to few or many tissues). The figure considers only PPIs associated with at least 1 tissue, and shows the percentage of these PPIs (y-axis) associated with up to a given percentage (i.e. $\leq k$ percent) of tissues (x-axis)

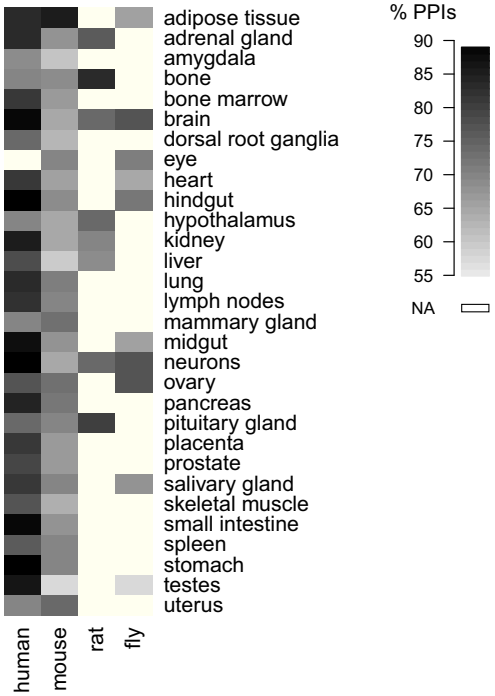

Figure 3. Tissue distribution of PPIs. Shown are percentages of each species' PPIs in given tissues.

Querying IID

Querying by protein or gene IDs. The main page of the IID website accepts gene or protein IDs and returns their PPIs. Inputs can be any combination of gene symbols, UniProt IDs or Entrez IDs separated by spaces, tabs or new lines. The IDs may be from one or more of the six species in IID. Checkboxes beside the input window control the types of evidence that are required for interactions: experimental detection, orthology or computational prediction. The second section of the page controls which species are considered in the search. Any combination of species can be selected from the list. Two checkboxes beside the list control how IID searches across species. One checkbox determines whether IID uses orthologs of input proteins in its search, if the proteins are not from the selected species. A second checkbox controls whether returned interactions can be in any of the selected species (default) or should be conserved across all selected species. The third section of the page controls which tissues will be considered in the search. Any combination can be selected from a list of 30 tissues, but not all tissues are

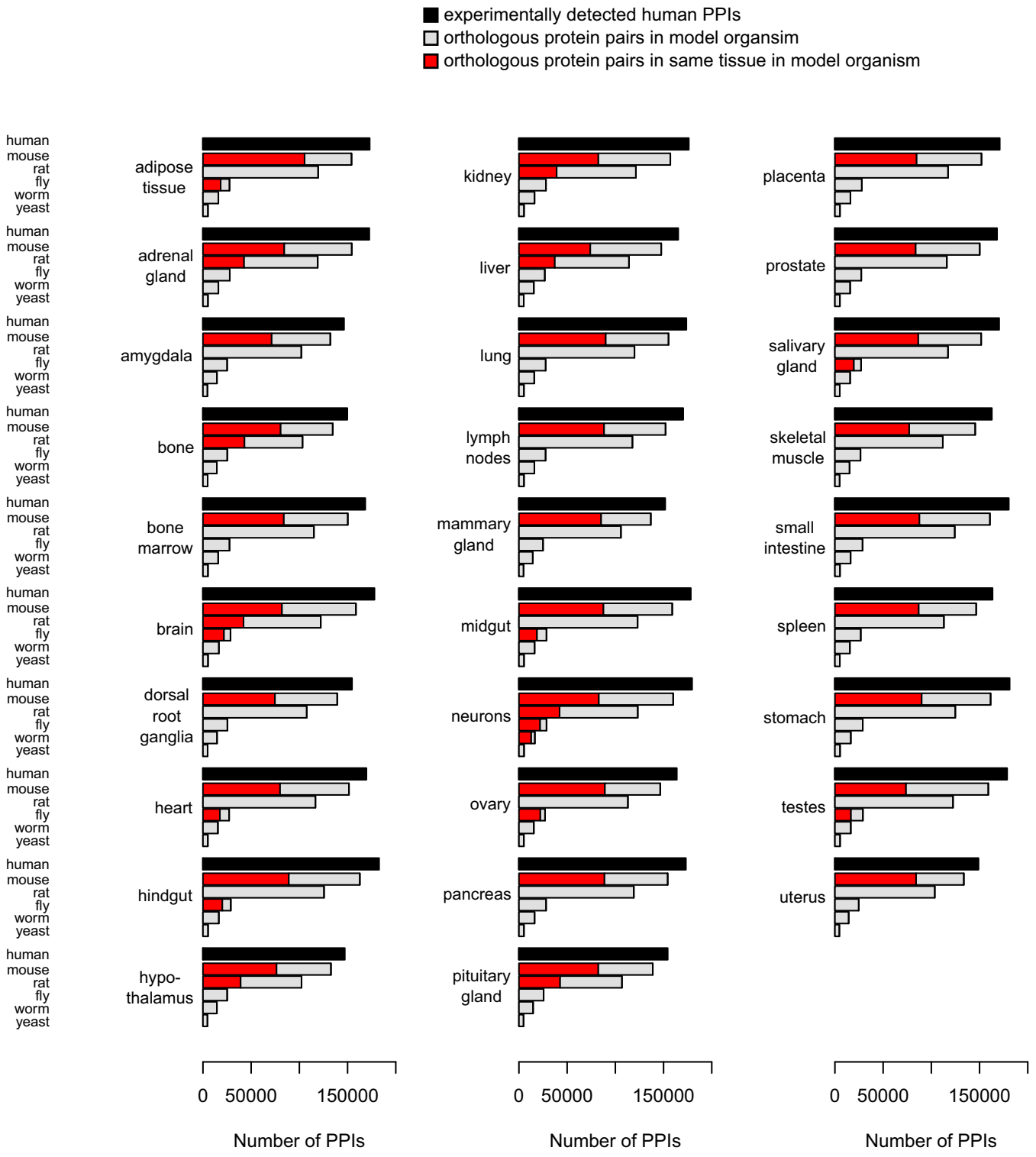

Figure 4. Conservation of human tissue-specific PPI networks in model organisms. IID annotates human PPIs with up to 29 tissues. For each of these tissues, the figure shows the number of experimentally detected human PPIs annotated with the tissue (black), the numbers of orthologous protein pairs in model organisms (grey) and the numbers of orthologous pairs annotated with the same tissue in the model organisms (red).

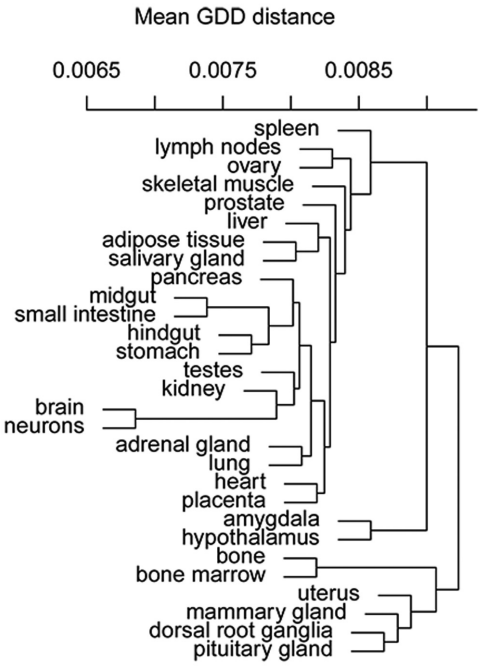

Figure 5. Clustering human tissues by graphlet degree distributions.

available in each species—a help symbol beside the tissue list shows available tissues for each species. Checkboxes beside the list control how IID searches across tissues: whether interactions can be in any selected tissues (default) or must be present in all, and whether tissue annotations can be based on either gene or protein expression, or must be based on both. If the selected tissue is set to 'any', IID does not filter interactions by tissue.

The last section of the page controls the output from IID. Search results can be displayed as a table, downloaded to a text file or viewed as a series of summary figures. A table format shows one interaction per row: two protein IDs, species and evidence. Users can choose to include information about the source of the interaction and the tissues where it is present. A graphical summary of results shows the percentages of interactions in different species and tissues, and network topology characteristics of each input protein. Network topology is analysed using graphlets (35) and displayed as graphlet degree distributions for each input protein. We found that clustering human tissue networks by graphlet degree distribution distance (35) identified expected similarities (e.g., amygdala and hypothalamus) and unforeseen ones (e.g., adipose tissue and salivary gland) (Figure 5).

Querying by interaction. Querying IID with a set of interactions can serve a number of useful functions: annotating interactions with evidence or tissues, filtering by evidence or tissues, mapping to orthologous interactions in other species, keeping only interactions conserved across species and many other possibilities. The input, selection, and output for this query are the same as for single proteins. Input is still a list of gene symbols, UniProt IDs or Entrez IDs, except IID assumes that every consecutive pair of IDs is an interaction. Query interactions can be mapped to other species simply by selecting these species from the list. Interactions can be filtered by selecting species or tissues; users can specify whether retained interactions should be in at least one of the selected species or tissues, or in all of them.

DISCUSSION

IID is the next generation of the I2D database, providing tissue-specific networks, new query and visualization capabilities and 74% more interactions. Its tissue annotations are available for human and four model organism networks, and are based on gene expression and proteomics data; an interaction is assumed to occur in a tissue if the two proteins or encoding genes are expressed in the tissue. IID allows users to easily find tissue-specific interactions of their proteins across multiple species—with the option of retaining only interactions conserved across species or tissues. IID also provides queries by interactions, allowing users to quickly annotate their network with interaction evidence or tissues, filter by evidence, tissues, or species, and map their network to other species. To provide more comprehensive networks, IID includes PPI predictions from four independent studies (23–26), totalling 664,643 interactions.

IID's method of mapping interactions to tissues, while commonly used (8–11), does not guarantee that an interaction will occur in a tissue. For example, expression of two genes in a given tissue may not mean that their two proteins will be present as well (36). Even if the two proteins are present, an interaction may not occur due to numerous reasons such as inappropriate sub-cellular localizations or post-translational modifications. Conversely, when an interaction is not mapped to a tissue, the interaction may still occur in the tissue under certain conditions. IID tissue assignments only indicate increased or decreased chances of occurrence.

Despite this uncertainty, tissue annotations still provide key benefits. For most applications of PPI networks it is essential to separate interactions that are happening in one tissue and not in another; otherwise the network may have little relation to the tissues being studied. For example, a cardiologist would need interactions typically present in heart tissue, and would need to exclude interactions that only occur in other tissues. In other cases, for example when testing a drug in a mouse model of human disease, it is more important to consider interactions that are shared between organisms and/or tissues. Applications of IID include selection of animal models, drug target discovery and pathway redefinition.

IID will be continuously maintained and updated every 6 months. Moreover, a curation of disease related interactions will be performed to include more specificity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Richard Lu and Mark Abovsky who maintain I2D.

FUNDING

University of Toronto McLaughlin Centre; Natural Sciences Research Council [NSERC 203475]; Canada Foundation for Innovation [CFI 12301, 203373, 29272, 225404, 30865]; Canada Research Chair Program [CRC 203373, 225404]; Ontario Research Fund [RE-03-020]; Ontario Research Fund [GL2-01-030]; US Army DOD W81XWH-12-1-0501; IBM and Ian Lawson van Toch Fellowship Award. Funding for open access charge: Ontario Research Fund [GL2-01-030]; Canada Research Chair Program [CRC 203373, 225404].

Conflict of interest statement. None declared.

REFERENCES

1. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
2. Mostafavi,S. and Morris,Q. (2012) Combining many interaction networks to predict gene function and analyze gene lists. *Proteomics*, **12**, 1687–1696.
3. Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057–1063.
4. Wang,X., Gulbahce,N. and Yu,H. (2011) Network-based methods for human disease gene prediction. *Br. Funct. Genomics*, **10**, 280–293.
5. Barabasi,A.L., Gulbahce,N. and Loscalzo,J. (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.
6. De Las Rivas,J. and Prieto,C. (2012) Protein interactions: mapping interactome networks to support drug target discovery and selection. *Methods Mol. Biol.*, **910**, 279–296.
7. Schaefer,M.H., Lopes,T.J.S., Mah,N., Shoemaker,J.E., Matsuoka,Y., Fontaine,J.-F., Louis-Jeune,C., Eisfeld,A.J., Neumann,G., Perez-Iratxeta,C. *et al.* (2013) Adding protein context to the human protein–protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.*, **9**, e1002860.
8. Barshir,R., Basha,O., Eluk,A., Smoly,I.Y., Lan,A. and Yeger-Lotem,E. (2013) The TissueNet database of human tissue protein–protein interactions. *Nucleic Acids Res.*, **41**, D841–D844.
9. Bossi,A. and Lehner,B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.*, **5**, 260.
10. Lopes,T.J.S., Schaefer,M., Shoemaker,J., Matsuoka,Y., Fontaine,J.-F., Neumann,G., Andrade-Navarro,M.A., Kawaoka,Y. and Kitano,H. (2011) Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, **27**, 2414–2421.
11. Liu,W., Wang,J., Wang,T. and Xie,H. (2014) Construction and analyses of human large-scale tissue specific networks. *PLoS One*, **9**, e115074.
12. Veres,D.V., Gyurkó,D.M., Thaler,B., Szalay,K.Z., Fazekas,D., Korcsmáros,T. and Csermely,P. (2015) ComPPI: a cellular compartment-specific database for protein–protein interaction network analysis. *Nucleic Acids Res.*, **43**, D485–D493.
13. Stumpf,M.P., Thorne,T., de Silva,E., Stewart,R., An,H.J., Lappe,M. and Wiuf,C. (2008) Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 6959–6964.
14. Chatr-Aryamontri,A., Breitkreutz,B.-J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
15. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
16. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
17. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
18. Breuer,K., Foroushani,A.K., Laird,M.R., Chen,C., Sribnaia,A., Lo,R., Winsor,G.L., Hancock,R.E.W., Brinkman,F.S.L. and Lynn,D.J. (2013) InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res.*, **41**, D1228–D1233.
19. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
20. Turinsky,A.L., Razick,S., Turner,B., Donaldson,I.M. and Wodak,S.J. (2014) Navigating the global protein–protein interaction landscape using iRefWeb. *Methods Mol. Biol.*, **1091**, 315–331.
21. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
22. Kamburov,A., Stelzl,U., Lehrach,H. and Herwig,R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793–D800.
23. Elefsinioti,A., Saraç,Ö.S., Hegele,A., Plake,C., Hubner,N.C., Poser,I., Sarov,M., Hyman,A., Mann,M., Schroeder,M. *et al.* (2011) Large-scale de novo prediction of physical protein–protein association. *Mol. Cell. Proteomics*, **10**, M111.010629.
24. Zhang,Q.C., Petrey,D., Deng,L., Qiang,L., Shi,Y., Thu,C.A., Bisikirska,B., Lefebvre,C., Accili,D., Hunter,T. *et al.* (2012) Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.
25. Kotlyar,M., Pastrello,C., Pivetta,F., Lo Sardo,A., Cumbaa,C., Li,H., Naranian,T., Niu,Y., Ding,Z., Vafaee,F. *et al.* (2015) In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat. Methods*, **12**, 79–84.
26. Rhodes,D.R., Tomlins,S.A., Varambally,S., Mahavisno,V., Barrette,T., Kalyana-Sundaram,S., Ghosh,D., Pandey,A. and Chinnaiyan,A.M. (2005) Probabilistic model of the human protein–protein interaction network. *Nat. Biotechnol.*, **23**, 951–959.
27. Brown,K.R. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, R95.
28. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
29. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
30. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419–1260419.
31. Wang,M., Herrmann,C.J., Simonovic,M., Szklarczyk,D. and von Mering,C. (2015) Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, **15**, 3163–3168.
32. Database resources of the National Center for Biotechnology Information. (2014) *Nucleic Acids Res.*, **42**, D7–D17.
33. The UniProt Consortium. (2014) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
34. Hočevar,T. and Demšar,J. (2014) A combinatorial approach to graphlet counting. *Bioinformatics*, **30**, 559–565.
35. Przulj,N. (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, e177–e183.
36. Vogel,C. and Marcotte,E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, **13**, 227–232.