Published online 29 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D419–D426***
doi: 10.1093/nar/gkw1082

pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis

Sara Rahmati[1], Mark Abovsky[2], Chiara Pastrello[2] and Igor Jurisica[1,2,3,4,*]

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada, [2]Princess Margaret Cancer Centre, University Health Network, 101 College Street, TMDT, Room 11–314, Toronto, ON M5G 1L7, Canada, [3]Department of Computer Science, University of Toronto, Toronto, ON, Canada and [4]Institute of Neuroimmunology, Slovak Academy of Sciences, Bratislava, Slovakia

Received August 15, 2016; Revised September 30, 2016; Editorial Decision October 24, 2016; Accepted October 25, 2016

ABSTRACT

Molecular pathway data are essential in current computational and systems biology research. While there are many primary and integrated pathway databases, several challenges remain, including low proteome coverage (57%), low overlap across different databases, unavailability of direct information about underlying physical connectivity of pathway members, and high fraction of protein-coding genes without any pathway annotations, i.e. 'pathway orphans'. In order to address all these challenges, we developed pathDIP, which integrates data from 20 source pathway databases, 'core pathways', with physical protein–protein interactions to predict biologically relevant protein–pathway associations, referred to as 'extended pathways'. Cross-validation determined 71% recovery rate of our predictions. Data integration and predictions increase coverage of pathway annotations for protein-coding genes to 86%, and provide novel annotations for 5732 pathway orphans. PathDIP (http://ophid.utoronto.ca/pathdip) annotates 17 070 protein-coding genes with 4678 pathways, and provides multiple query, analysis and output options.

INTRODUCTION

Together with other biological networks physical protein–protein interaction (PPI) networks and pathways are essential resources in computational and systems biology research (1–3). Known and predicted PPIs connect 17 976 human proteins by 850 636 interactions (4), yet this number will increase in the future due to more splice variants-(5) and context-specific PPIs (6). 'Contextualizing' the networks, by identifying conditions such as localization, cell type, tissues and processes where these interactions are functional is a multi-step process. Some of this information is available in existing source pathway databases, which provide detailed information about functional interactions of biomolecules, including proteins, that work cooperatively to accomplish a specific task (7).

Primary pathway databases provide manually curated resource for pathway models, focusing either on specific processes (e.g. NetPath (8), SMPDB (9)), or global characterization (e.g. KEGG (10), Reactome (11)). Besides providing detailed information about molecular dynamics and cellular processes, these resources are essential for pathway annotation and enrichment analysis (reviewed in (7,12)).

Despite these efforts, analysing data through primary pathway databases remains challenging due to: (i) low protein-coding gene coverage of individual databases that significantly biases analysis (e.g. KEGG and Reactome cover 6724 and 7667 unique protein-coding genes, respectively), (ii) low overlap among different databases that leads to different enrichment analysis results (KEGG and Reactome overlap in 4992 genes while their union covers 9396 genes), (iii) lack of information about physical vs functional interactions, resulting in missing key information and (iv) high fraction of protein-coding genes being absent from any database, referred to as '*pathway orphans*'.

Integrated pathway databases such as DAVID (13), iPavs (14), Panther (15), PathCards (16), PathwayCommons (17) and WikiPathways (18), address low overlap of primary databases. ConsensusPathDB (19) and EnrichNet (20) attempt to reduce study biases by analysing network modules specific to each query gene list, and HCPIN (21) extracts pathway networks by focusing on structural analysis of cancer pathways, limiting its annotation domain. While these efforts increase coverage and improve enrichment analysis, they still suffer from pathway orphans.

To address these shortcomings, we integrated data of twenty core pathway databases with physical protein interactions to predict missing protein–pathway associa-

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

tions. We have developed pathway Data Integration Portal, pathDIP (http://ophid.utoronto.ca/pathdip), for comprehensive gene enrichment analysis, which annotates 17 070 human protein-coding genes with 4,678 pathways.

MATERIALS AND METHODS

Data collection and processing

All data sources used in pathDIP are publically available, as described below. Details about data sources versions and our processing methods are available in Supplementary Table S1A.

Pathways. PathDIP provides access to 4678 pathways from twenty source pathway databases (Supplementary Table S1B).

Protein interactions. Experimentally detected physical PPIs and high-confidence computationally predicted PPIs based on orthology and FpClass algorithm (22) were obtained from IID version 2015-09 (4).

Diseases. We integrated data from three disease gene databases:

1. GAD (23) includes 15 160 disease titles in 19 categories of diseases (final release: September 2014);
2. COSMIC (24) archives detected mutations in genes across 43 different tissues by curating low-throughput data from papers as well as processing high-throughput data from genome-wide studies, such as TCGA and ICGC. We defined and used COSMIC-05 as list of COSMIC genes with mutation rate of at least 5% in at least one tissue;
3. GeneSigDB (25) is a database of manually curated disease gene signatures extracted from published papers across different groups of diseases. GeneSigDB v.4 includes 2906 human cancer gene signatures from 1366 papers, covering 17 859 unique protein-coding genes.

Overlap of these three databases with pathDIP data is presented in Supplementary Table S2A.

Drugs. We obtained druggable genome list and drugs in clinical trial database from (26,27).

GeneOntology (*GO*). We downloaded human gene annotations file from GO website (http://geneontology.org) as of 12 October 2015.

ID conversion protocol

While pathDIP handles multiple IDs, e.g. Entrez gene, UniProt, Ensemble, gene symbols, etc., we use Entrez Gene ID as primary identifier in this paper.

Protein–pathway association scores

First, we defined each pathway, P, as a feature and tagged each protein in the PPI network with its corresponding features, i.e. pathways that it belongs to. Next, for each protein p we defined n_f as the number of its interaction neighbours that were tagged with feature f. Finally, we used Fisher's exact test to calculate the probability by which we expect p to have equal or more than n_f neighbours with tag f (details are provided in Supplementary text). We corrected these calculated probabilities for multiple hypothesis testing (FDR; BH function in R) to obtain $fdr_{(p,\,P)}$. We defined association score of a pair $(p,\,P)$ as:

$$assoc_{(p,\,P)} = 1 - fdr_{(p,\,P)}.$$

Protein–pathway pairs with score of at least 0.95 (i.e. pairs with FDR < 0.05) are predicted to be associated.

Evaluation of predictions

Data. We used extended pathways that predict novel pathway associations using integrated experimentally detected and computationally predicted PPIs described above. For analysis and statistics reported in this paper we used cut-off threshold of 0.95 for association score.

Randomization test. Since we cannot evaluate sensitivity directly (due to lack of true negatives), we evaluate the scale and significance of the recovery rate (details are provided in Supplementary text) of our predictions by comparing it to an empirical distribution of recovery rates for predictions that were obtained by using randomized networks. We randomly shuffled labels of proteins in union of interactome and pathways (details are provided in Supplementary text) 10 000 times, while maintaining the topology of PPI network and pathway annotations of each protein (i.e. genes with same labels in all randomized networks have same pathway annotations too). We executed our scoring algorithm using these randomly labelled PPI networks to generate 10 000 random-based extended pathways. Next, we compared the recovery rate as well as distribution of association scores obtained from these random-based extended pathways, with those that were obtained from original PPI network (whose protein labels are correct).

Overlap coefficient. We calculated overlap coefficient for each pathway pair (Supplementary Figures S1A and B) by dividing size of their overlap over size of the smaller pathway between the two pathways under comparison.

Jaccard index. We calculated Jaccard Index for each pathway pair (Supplementary Figures S1C and D) by dividing size of their overlap over size of union of the two pathways under comparison.

pathDIP portal

PathDIP provides users with several filters and search settings to customize the retrieval and analysis.

Customized search. PathDIP accepts lists of different types of identifiers. Users can search by selecting any combination of twenty source databases, core or extended pathways, and confidence level of the predictions.

Summary statistics. Coverage of each pathway set, i.e. core and four different extended pathway sets, for the input list is provided to users to facilitate more educated choice of settings when they customize their search.

Pathway association annotations. PathDIP annotates proteins in the input list with the predicted associated pathways. For each novel prediction, association score, and PPIs that connect protein and pathway are included in this table, along with links to relevant Entrez Gene, UniProt, and HGNC entries. Summary of the associations are also available in a matrix view. Both views are downloadable in tab-delimited format. Users can select one of the five pathway definition sets: core pathways, extended pathways based on experimentally detected PPIs at cut-off association score of 0.99 or 0.95, and extended pathways based on experimentally and high-confidence computationally predicted PPIs at cut-off association score of 0.99 or 0.95 as source of annotation and background for enrichment analysis.

Enrichment analysis. PathDIP portal provides pathway and pathway title word frequency enrichment analysis. In both cases, we use Fisher's exact test and correct raw *P*-values for multiple hypothesis testing based on two methods: Bonferroni and false discovery rate (BH method).

Pathway enrichment analysis. A detailed table comprising pathway enrichment analysis results is available in both HTML and downloadable tab-delimited formats. All of the genes that belong to the user-selected pathway definition sets (e.g. core or extended pathways) and source databases (e.g. all twenty databases or a subset of them) are considered as background to calculate row over-representation probabilities at a query time.

Word enrichment analysis. Depending on the input list and user settings, interpreting and visualizing pathway enrichment results can be challenging due to large result tables. PathDIP offers pathway title word enrichment analysis, where words in all pathway titles serve as background to calculate enrichment score for each informative word present in the titles of enriched pathways. Next, for each word W with enrichment score $P(_W)$, we defined a word enrichment score, $S(_W)$, as: $-log_{10}\ (P(w))$. While pathDIP provides simple word cloud based on word enrichment scores, words and their enrichment scores are also available for download as a text file, which could be visualized using customisable word cloud visualizer such as http://www.wordle.net, https://www.jasondavies.com/wordcloud/, and wordcloud package in R. Resulting word clouds do not represent simple word frequencies, rather, they reflect the significance of appearance of each word in titles of enriched pathways.

Programming tools

Predictions and analysis. Association scores were computed in R (version 3.0.2), and its base packages and functions. Graphs in this paper were generated using RColorBrewer (version: 1.1_2), gplots (version: 2.16.0), VennDiagram (version: 1.6.9), ComplexHeatmap (version: 1.11.1), and circlize (version: 0.3.7) packages.

Table 1. Recovery rate of predictions

Recovery	Only experimental PPIs	Experimental and high-confidence predicted PPIs
assoc$(g,P) \geq 0.99$	0.51	0.65
assoc$(g,P) \geq 0.95$	0.60	0.71

Portal implementation. Interface to the pathDIP database is implemented at JavaServer Faces framework running on IBM WebSphere application server (v8.5), with IBM DB2 database (v10.1) engine as a back end storage. In order to improve the performance, the WebSphere and DB2 are running on different virtual instances of IBM P770 and P750, running AIX (v7.1).

RESULTS AND ANALYSIS

PathDIP content

Pairwise overlap of existing primary pathway databases remains limited—only 4992 protein-coding genes overlap between the largest databases (KEGG and Reactome; Figure 1A and B), and only four genes overlap between BioCarta and SIGNOR. While this causes inconsistent results of pathway enrichment analysis across different databases, it shows that large fraction of data is complementary (Figure 1B and C), and would benefit from integration.

Core pathways. Core pathways integrate 4678 pathways from twenty pathway sources (almost three times larger than Reactome), and annotate 11 338 human unique protein-coding genes (providing 57% protein-coding gene coverage). Gene coverage saturates at 10 605 after integrating the six largest source pathway databases (Supplementary Figure S2) and 43% of human protein-coding genes remain pathway orphans, i.e. proteins lacking pathway annotation in any of the twenty databases (Figure 1D), which leads to substantial biases in pathway enrichment analysis.

Pathway associations and their computational evaluation. We used physical PPIs to predict biologically relevant protein–pathway associations. Computational evaluation of our predictions for core protein–pathway pairs shows that: (i) distribution of association scores for core protein–pathways is significantly higher than that of the full set of association scores (Supplementary Figures S3A and B); (ii) recovery rate of our predictions for core protein–pathway pairs is 71% (Table 1), 24 times larger than recovery rate of predictions based on random networks (P-value $< 10^{-4}$; Supplementary Figures S3C–H).

Extended pathways. Extended pathways integrate core pathways and predicted protein–pathway associations, and annotate 17 070 protein-coding genes (Figure 2A and B), which is over 2.5 and 2.2 times larger than coverage of KEGG and Reactome, individually. Extended pathDIP also increases pairwise overlap of pathway databases (Figures 2C–E and Supplementary Figures S1A and B). Importantly, this improvement is not a result of associating all genes to all pathways (Supplementary Figures S1B and D).

Increased coverage significantly contributes to more systematic analysis and reduces bias caused by lack of path-

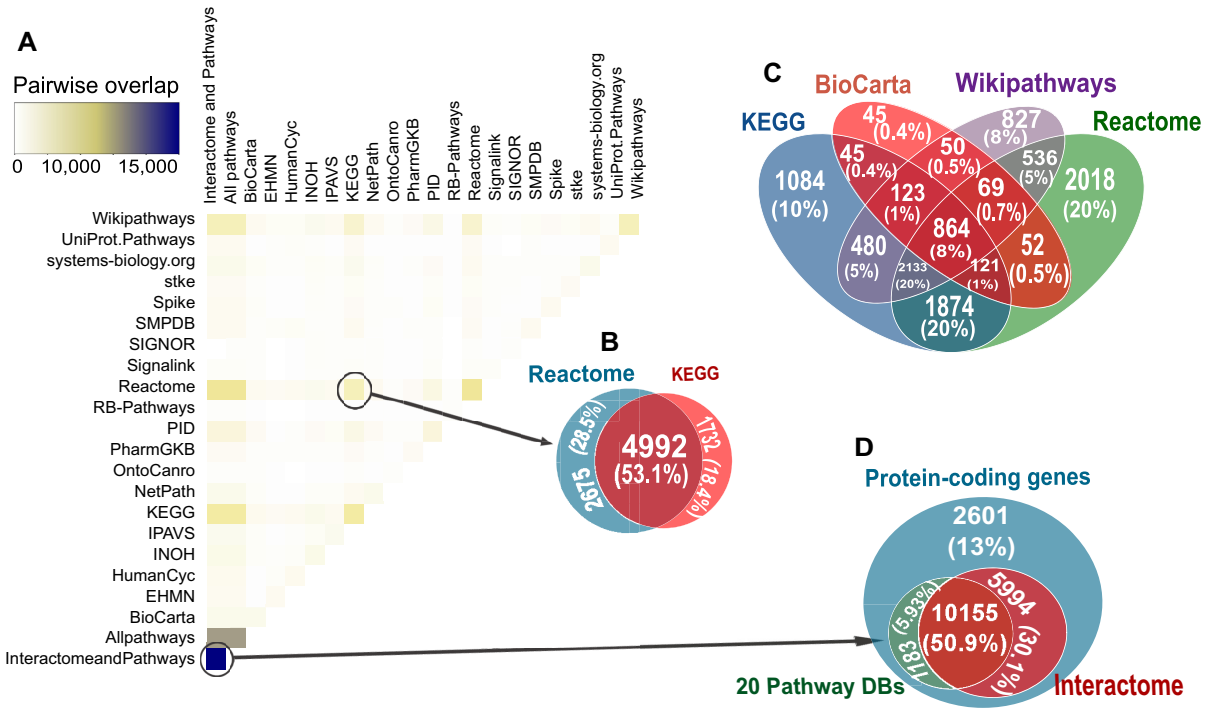

Figure 1. Coverage of major pathway databases and physical PPIs for protein-coding genes. (**A**) Number of proteins covered by individual pathway databases (diagonal), and their pairwise overlaps for protein-coding genes. (**B**) Each of the two largest core pathway databases covers less than 40% of protein-coding genes. Moreover, <25% of the genes ($n = 4992$) are covered by both KEGG and Reactome. (**C**) Overlap of the four largest available general-domain pathway databases shows that they share only 864 genes (8.4% of the union of four pathway databases). (**D**) 43% of protein-coding genes do not have any pathway annotation, from which 5994 are available in interactome.

way annotation. For example, while each of KEGG, Reactome, and core pathDIP cover between 0% and 100% of each of the 2906 human cancer gene signatures in GeneSigDB (25) median coverage of KEGG and Reactome are 55% and 57%, while median coverage of core pathDIP is 81%. Minimum coverage of gene signatures by extended pathDIP is increased to 59% and its median reaches 99% (boxplots in Figure 3A). Furthermore, minimum coverage of extended KEGG and extended Reactome is increased to 40% and 47% with median coverage of 95% and 97% (details in Supplementary Table S2B).

Core and Extended pathways are available for download and enrichment analysis can be performed online at pathDIP portal: http://ophid.utoronto.ca/pathdip.

DISCUSSION

PathDIP is a pathway annotation database and enrichment analysis portal that associates 17 070 unique human protein-coding genes to 4678 pathways. Notably, pathDIP provides pathway annotations for 5732 pathway orphans. Absence of these genes from core pathways is not due to their lack of importance as many of them are known disease genes and drug targets. For example, 'druggable genome' list (26) includes 3860 unique genes out of which 937 are pathway orphans, and 560 of them are annotated in extended pathDIP (Supplementary Tables S2C–E). Our analysis also shows that many of pathway orphan genes have been listed as disease genes in different disease databases. We found 503 pathway orphans that belong to three major disease gene databases, GeneSigDB, GAD and COSMIC-05, and extended pathDIP annotates 487 of those orphans (Figure 3B, Table 2, and Supplementary Tables S2F–G).

While all 487 disease genes are important, *PTPRD* has been associated to the highest number of diseases in GAD (54 diseases and 13 disease classes). It has been mutated in >13% of melanoma samples, 9% of lung adenocarcinomas (7% of all sub-types of lung cancer), and 8% of stomach cancers based on COSMIC data (Supplementary Tables S3A–D). *PTPRD* is also among genes in druggable genome. A large list of publications that support the importance of *PTPRD* in diseases is reviewed in (28). Our predictions have associated this gene to 213 different pathways (Supplementary Table S3E). 'Signaling', 'adherens', 'neurotransmitter', 'cancer', 'EGFR', 'ERBB2', 'MAPK', 'E-cadherin, 'SHC1', 'adhesion', are among top enriched words in titles of these pathways (Supplementary Table S3F). This is highly consistent with GAD classes of diseases that *PTPRD* has been associated with (Supplementary Table S3B), as well as with Gene Ontology terms that *PTPRD* is annotated with (Supplementary Table S3G).

Low coverage of existing primary and integrated pathway databases compared with Gene Ontology (GO) has resulted in frequent application of GO instead of pathways for gene/protein enrichment analysis, at the price of losing information about physical and functional dependencies among genes. Integrating pathway resources with PPIs addresses this challenge, since median coverage of extended pathDIP and GO-human for genes in 2906 human cancer gene signatures available in GeneSigDB (sizes range be-

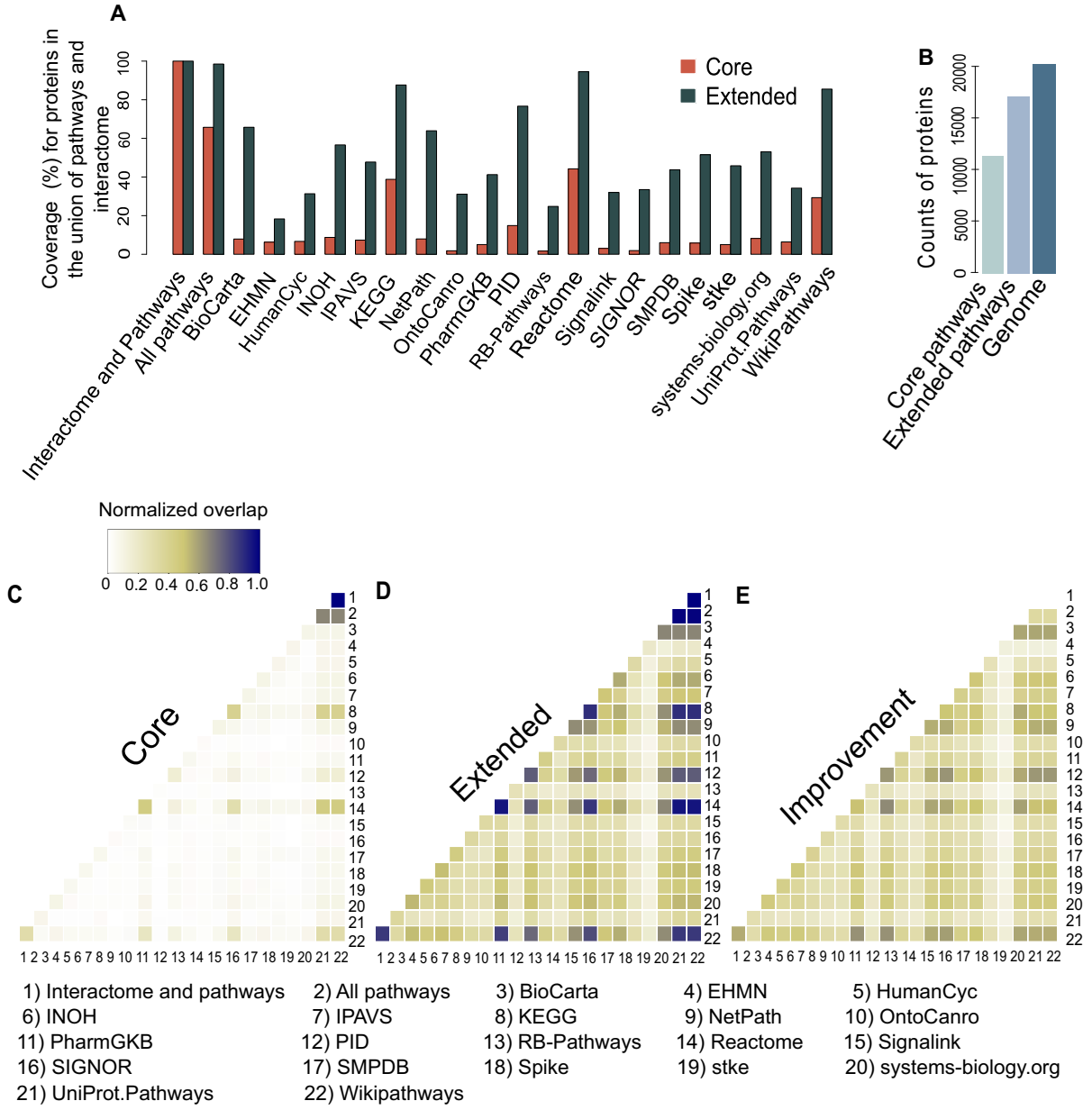

Figure 2. Improvement of core *vs* extended pathway databases in covering genes/proteins. (**A**) Comparison of coverage of core versus extended version of each pathway database for interactome. (**B**) Overall coverage of core and extended pathways for protein-coding genes. (**C**) Pairwise overlap of core pathway databases. (**D**) Pairwise overlap of extended pathway databases. (**E**) Extended pathways show significant improvement in the protein coverage of single databases (diagonal) and pairwise overlaps of pathway pairs (under the diagonal), compared with core databases.

Table 2. Pathway orphan genes associated to diseases in different disease databases and their overlaps

Disease databases	GeneSigDB	GAD	COSMIC-05	GeneSigDB, GAD	GeneSigDB, COSMIC-05	GAD, COSMIC-05	GeneSigDB, GAD, COSMIC-05
Number of pathway orphans with available predictions	2349	12	10	2537	165	26	478

tween 1 to 3927 genes with median of 37 and mean of 112) reaches 99%. Importantly, although distributions of pathDIP and GO coverage for cancer signatures are comparable (P-value < 0.45, Student's t-test) minimum coverage of GO is 41% but pathDIP covers at least 59% of each gene signature (Figure 3A). Notably, comparison of any other two out of these five functional annotation sets is statistically significant (Figure 3A). Other advantages of pathDIP over GO for enrichment analysis include providing list of physical interactions of protein–pathways that is missing in GO, as well as a smaller set of terms (about 1/10 of GO) that are also more informative and specific.

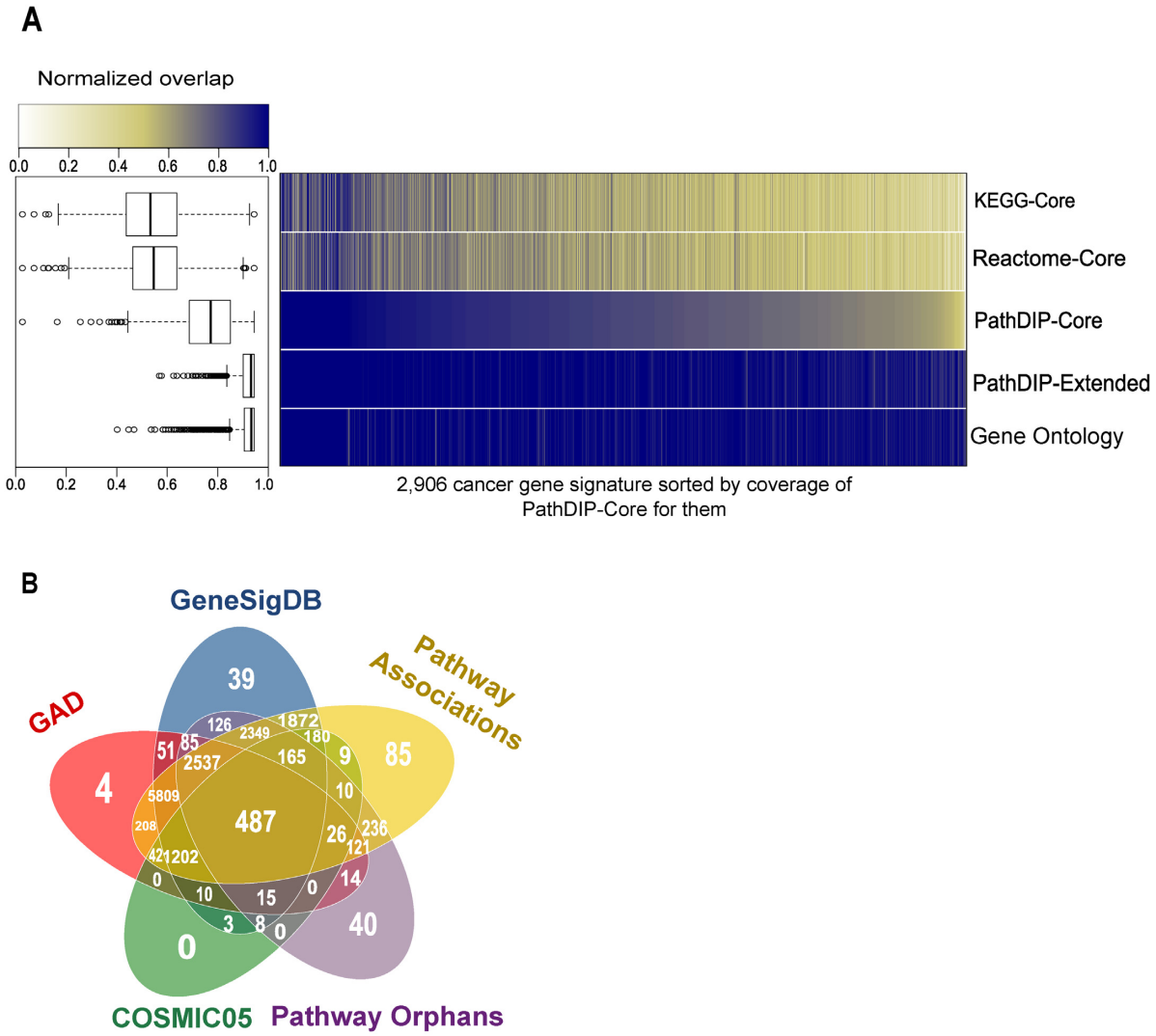

Figure 3. Coverage of pathDIP for disease genes and its comparison with other databases. (**A**) Normalized coverage of different functional annotation databases for cancer genes shows that extended pathDIP is more comprehensive than other pathway databases and is comparable with GO. Furthermore, while pairwise comparison of coverage of KEGG, Reactome, Core pathDIP with each other and with both extended pathDIP and GO for signatures shows statistically significant (two-tailed Student's *t*-test) distributions (KEGG and Reactome: $P < 0.005$, KEGG/Reactome and Core pathDIP: $P < 2.2 \times 10^{-16}$, Core pathDIP and Extended pathDIP/GO: $P < 2.2 \times 10^{-16}$), GO and Extended pathDIP do not show any significant difference in distributions of their coverages for gene signatures (P-value < 0.45). Boxplots show that although coverages of pathDIP and GO for cancer signatures is similar, the minimum coverage by GO is 41%, while pathDIP covers at least 59% of each cancer gene signature. (**B**) Overlap of GeneSigDB, GAD and COSMIC covers 1716 genes out of which 503 are pathway orphans and we have annotated 487 of them with predicted pathway associations. Furthermore, there are several thousands of pathway orphan genes that have been associated with diseases in at least two out of three databases for which pathway associations are available.

PathDIP not only increases coverage of pathway annotations for proteins, it also improves pairwise and overall overlaps of pathways of the same-class. Same-class pathways are pathways from different source databases that define the same process or function. Supplementary Figures S4 and S5 compare distribution of overlap of pathway pairs across and within pathway databases, and demonstrate that low overlap of pathway databases is relatively evenly distributed across pathway pairs regardless of whether they are the same-class or not (see Supplementary text and Supplementary Table S4 for more details). To illustrate this improved feature in pathDIP, we chose five of the most studied pathway classes among hallmarks of cancer (29): 'cell cycle', 'apoptosis', 'NOTCH1 signalling', 'P53 signalling' (four pathways frequently disrupted across different cancer types (29,30)), and 'glucose metabolism' (at top of hierarchy of cancer metabolic pathways (29,31)). Overlaps of extended definitions of pathways in each of these classes exhibit clear improvement compared with the overlaps of their core members (Supplementary Figures S6 and S7). Increased overlap of same-class pathways also implies that, although pathways in the same class share small number of core members, their member genes share significantly high number of direct interactors in PPI network that results in associating same proteins to same-class pathways. Glucose

metabolism pathway-class is the only exception among the above five pathway-classes.

Our analysis shows that metabolic pathways in general, are one group of pathways with almost no expansion. In fact, extended pathways' growth rate compared with their core versions is not uniformly distributed. While the median of growth rate of pathway sizes is 9.88 (Supplementary Table S5A), some pathways show growth rate above 500 times and some show close to zero growth rate. The top 5% extended pathways include signalling, cancer, and DNA damage and repair pathways (Supplementary Table S5B). High growth rate of these pathways is due to presence of highly studied genes such as *MAPK*, *PIK3*, *TP53*, *GRB2*, *HRAS*, *AKT*, *EGFR*, etc. in these pathways. The bottom 5% extended pathways include olfactory, metabolism and biosynthesis, transporters, and receptors pathways [Supplementary Table S5C]. Low growth rate of metabolic pathways is also reflected in Figure 2E in which EHMN and HumanCyc, the two metabolic pathway databases, show low improvement in their protein coverage. We attribute the low growth rate of these pathways to low degree proteins or interactome orphans (22). This is consistent with the results of (22) that also suggests reasons for such proteins to lack PPIs in currently available interactome.

This direct effect of PPI networks on coverage of pathDIP exemplifies the fact that, despite its unique features, extended pathDIP is still bound to available biases in its data sources, i.e. pathways and PPI data. In particular, extended pathDIP cannot annotate a few thousands of protein-coding genes that are currently absent from PPI network, many of which are important in medicine. For example, among 937 pathway orphan genes in druggable genome list, 377 genes are not present in extended pathDIP. Moreover, out of 1049 unique proteins that are annotated as drug targets in Drugs in Clinical Trials Database (27), 70 are pathway orphans. Twenty three of these proteins are targets of drugs some of which have been approved by FDA since 1956. PathDIP provides predicted pathway associations for only 15 of these 70 genes. The remaining 55 genes as well as 361 of the aforementioned 377 druggable genes are PPI orphans, i.e. they are absent from PPI network (22). Gene Ontology enrichment analysis of these proteins suggests their involvement in 'metal ion transmembrane transporter activity', 'transmembrane receptor activity', 'peptide binding', and 'cation channel activity', as well as 'protein metabolic process', 'cell-cell signalling' and 'response to stimulus'. Furthermore, there are 262 unique pathway orphan genes with protein products in the PPI network for which we could not make any predictions. These genes are of low degree in the PPI network, i.e. only 83 of them have degree >1 and the maximum degree is 14 (Supplementary Table S6). Low degree proteins in PPI network have intrinsic properties similar to properties of PPI orphan genes (22). Since these 262 pathway orphan genes are also low degree proteins in PPI network, their lack of pathway annotations could be due to such properties that make them more challenging to study.

Sources, analysis and results in this paper are based on pathDIP version 2.5, which has been built based on Fall/Winter 2015 releases of source pathways and PPI sources. Some of the pathway and interaction resources have been updated since then. For example, the most recent release of Reactome (v58, released on September 28, 2016) covers about 10,000 protein-coding genes, and KEGG has added or revised seventeen new human pathways to its database (Release 80.0, October 1, 2016). These and other improvements will be integrated into the next pathDIP release, as we plan two major releases annually.

Integrating curated pathways with PPI data and computational pathway-association predictions provides the most sensible approach to decreasing biases and improving pathway annotation coverage for protein-coding genes. PathDIP (http://ophid.utoronto.ca/pathdip) offers such rich resource for pathway annotation and pathway enrichment analysis.

FUTURE WORK

This manuscript and current pathDIP provide only the first step in addressing pathway integration and prediction challenges. One of the major remaining challenges is a comprehensive same-class core pathway consolidation across multiple sources. This is not a trivial task due to several problems including (but not limited to) different criteria that individual databases cover such as domain-specificity and confidence levels of pathway members and PPIs, as well as use of diverse vocabulary and terminologies in titles of similar pathways. For example, spliceosome pathway in KEGG is referred to as mRNA processing and in Reactome as mRNA splicing pathway. A related problem is redundancy in integrated pathway databases due to similar information provided by different pathway databases. A third problem is different pathways with close titles; for example, snRNP and vRNP (dis)assembly. These problems will influence pathway association prediction and enrichment analysis, and thus any solution would need to be thoroughly validated by biological experiments and curation. Addressing such challenges is part of our plans for further improvement of pathDIP.

Current version of pathDIP has been already applied (32,33), highlighting the value of predicted pathway associations in pathway enrichment analysis. Furthermore, improved overlap of extended pathways compared with overlap of their core versions (Supplementary Figures S6 and S7) suggests that extended source-specific, same-class pathways converge to similar definitions, highlighting that pathDIP will contribute to consolidation of existing pathways. Interestingly, same-class pathways are extended based on their low-overlapping core definitions (gene content). As such, it provides the consolidated infrastructure for addressing the next challenges such as consolidating pathways and validating predictions experimentally.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr Max Kotlyar for the valuable discussions through development of pathDIP and Richard Lu for helping with maintenance of database.

FUNDING

Ontario Research Fund [GL2-01-030] (in part); Canada Research Chair Program [CRC 203373, 225404]; Ontario Research Fund [RE-03-020]; Natural Sciences Research Council [NSERC 203475]; Canada Foundation for Innovation [CFI 12301, 203373, 29272, 225404, 30865]; US Army [DOD W81XWH- 1-1-0501]; IBM. Funding for open access charge: Canada Research Chair Program [CRC 225404].

Conflict of interest statement. None declared.

REFERENCES

1. Chuang,H.-Y., Lee,E., Liu,Y.-T., Lee,D. and Ideker,T. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.
2. Pržulj,N., Wigle,D. and Jurisica,I. (2004) Functional topology in a network of protein interactions. *Bioinformatics*, **20**, 340–348.
3. Bild,A.H., Yao,G., Chang,J.T., Wang,Q., Potti,A., Chasse,D., Joshi,M.B., Harpole,D., Lancaster,J.M., Berchuck,A. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.
4. Kotlyar,M., Pastrello,C., Sheahan,N. and Jurisica,I. (2016) Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.*, **44**, D536–D541.
5. Yang,X., Coulombe-Huntington,J., Kang,S., Sheynkman,G.M., Hao,T., Richardson,A., Sun,S., Yang,F., Shen,Y.A., Murray,R.R. *et al.* (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**, 805–817.
6. Snider,J., Kotlyar,M., Saraon,P., Yao,Z., Jurisica,I. and Stagljar,I. (2015) Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.*, **11**, 848–848.
7. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
8. Kandasamy,K., Mohan,S., Raju,R., Keerthikumar,S., Kumar,G.S.S., Venugopal,A.K., Telikicherla,D., Navarro,D.J., Mathivanan,S., Pecquet,C. *et al.* (2010) NetPath: a public resource of curated signal transduction pathways. *Genome Biol.*, **11**, R3.
9. Jewison,T., Su,Y., Disfany,F.M., Liang,Y., Knox,C., MacIejewski,A., Poelzer,J., Huynh,J., Zhou,Y., Arndt,D. *et al.* (2014) SMPDB 2.0: Big improvements to the small molecule pathway database. *Nucleic Acids Res.*, **42**, D478–D484.
10. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
11. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F. and McKay,S. (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
12. Xie,C., Mao,X., Huang,J., Ding,Y., Wu,J., Dong,S., Kong,L., Gao,G., Li,C.-Y. and Wei,L. (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.*, **39**, W316–W322.
13. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.*, **4**, 44–57.
14. Sreenivasaiah,P.K., Rani,S., Cayetano,J., Arul,N. and Kim,D.H. (2012) IPAVS: Integrated pathway resources, analysis and visualization system. *Nucleic Acids Res.*, **40**, D803–D808.
15. Mi,H., Poudel,S., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.
16. Belinky,F., Nativ,N., Stelzer,G., Zimmerman,S., Iny Stein,T., Safran,M. and Lancet,D. (2015) PathCards: Multi-source consolidation of human biological pathways. *Database: The Journal of Biological Databases and Curation*, **2015**, bav006.
17. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,O., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
18. Kutmon,M., Riutta,A., Nunes,N., Hanspers,K., Willighagen,E.L., Bohler,A., Mélius,J., Waagmeester,A., Sinha,S.R., Miller,R. *et al.* (2015) WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.*, doi:10.1093/nar/gkv1024.
19. Herwig,R., Hardt,C., Lienhard,M. and Kamburov,A. (2016) Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat Protoc.*, **11**, 1889–1907.
20. Glaab,E., Baudot,A., Krasnogor,N., Schneider,R. and Valencia,A. (2012) EnrichNet: Network-based gene set enrichment analysis. *Bioinformatics*, **28**, i451–i457.
21. Huang,Y.J., Hang,D., Lu,L.J., Tong,L., Gerstein,M.B. and Montelione,G.T. (2008) Targeting the human cancer pathway protein interaction network by structural genomics. *Mol. Cell Proteomics*, **7**, 2048–2060.
22. Kotlyar,M., Pastrello,C., Pivetta,F., Lo Sardo,A., Cumbaa,C., Li,H., Naranian,T., Niu,Y., Ding,Z., Vafaee,F. *et al.* (2014) In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat. Methods*, **12**, 79–84.
23. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
24. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
25. Culhane,A.C., Schröder,M.S., Sultana,R., Picard,S.C., Martinelli,E.N., Kelly,C., Haibe-Kains,B., Kapushesky,M., St Pierre,A.A., Flahive,W. *et al.* (2012) GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.*, **40**, D1060–D1066.
26. Wagner,A.H., Coffman,A.C., Ainscough,B.J., Spies,N.C., Skidmore,Z.L., Campbell,K.M., Krysiak,K., Pan,D., McMichael,J.F., Eldred,J.M. *et al.* (2016) DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.*, **44**, D1036–D1044.
27. Rask-Andersen,M., Masuram,S. and Schiöth,H.B. (2014) The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu. Rev. Pharmacol. Toxicol.*, **54**, 9–26.
28. Zhao,S., Sedwick,D. and Wang,Z. (2015) Genetic alterations of protein tyrosine phosphatases in human cancers. *Oncogene*, **34**, 3885–3894.
29. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
30. Leiserson,M.D.M., Vandin,F., Wu,H.-T., Dobson,J.R., Eldridge,J. V, Thomas,J.L., Papoutsaki,A., Kim,Y., Niu,B., McLellan,M. *et al.* (2014) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
31. Boroughs,L.K. and DeBerardinis,R.J. (2015) Metabolic pathways promoting cancer cell survival and growth. *Nat. Cell Biol.*, **17**, 351–359.
32. Li,Y.H., Tavallaee,G., Tokar,T., Nakamura,A., Sundararajan,K., Weston,A., Sharma,A., Mahomed,N.N., Gandhi,R., Jurisica,I. and Kapoor,M. (2016) Identification of synovial fluid microRNA signature in knee osteoarthritis: differentiating early- and late-stage knee osteoarthritis. *Osteoarthr. Cartil.*, **24**, 1577–1586.
33. Nakamura,A., Rampersaud,Y.R., Sharma,A., Lewis,S.J., Wu,B., Datta,P., Sundararajan,K., Endisha,H., Rossomacha,E., Rockel,J.S. 2016 *et al.* (2016) Identification of microRNA-181a-5p and microRNA-4454 as mediators of facet cartilage degeneration. *JCI Insight*, **1**, e86820.