

Reactome: a database of reactions, pathways and biological processes

David Croft¹, Gavin O'Kelly¹, Guanming Wu², Robin Haw², Marc Gillespie³, Lisa Matthews⁴, Michael Caudy², Phani Garapati¹, Gopal Gopinath⁵, Bijay Jassal¹, Steven Jupe¹, Irina Kalatskaya², Shahana Mahajan^{4,6}, Bruce May², Nelson Ndegwa¹, Esther Schmidt¹, Veronica Shamovsky⁴, Christina Yung², Ewan Birney¹, Henning Hermjakob¹, Peter D'Eustachio⁴ and Lincoln Stein^{2,7,8,*}

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK,

²Ontario Institute for Cancer Research, Informatics and Bio-Computing, Toronto, ON, M5G0A3, Canada,

³St. John's University, College of Pharmacy and Allied Health Professions, Queens, NY 11439, ⁴NYU School of Medicine, Department of Biochemistry, New York, NY 10016, ⁵US Food and Drug Administration, Center for Food Safety and Applied Nutrition, Laurel, MD 20708, ⁶Hunter College, School of Health Sciences, New York, NY 10010, ⁷Cold Spring Harbor Laboratory, Bioinformatics and Genomics, Cold Spring Harbor, NY 11724, USA, ⁸University of Toronto, Department of Molecular Genetics, Toronto, ON, M5S 1A1, Canada

Received September 15, 2010; Accepted October 10, 2010

ABSTRACT

Reactome (<http://www.reactome.org>) is a collaboration among groups at the Ontario Institute for Cancer Research, Cold Spring Harbor Laboratory, New York University School of Medicine and The European Bioinformatics Institute, to develop an open source curated bioinformatics database of human pathways and reactions. Recently, we developed a new web site with improved tools for pathway browsing and data analysis. The Pathway Browser is an Systems Biology Graphical Notation (SBGN)-based visualization system that supports zooming, scrolling and event highlighting. It exploits PSIQUIC web services to overlay our curated pathways with molecular interaction data from the Reactome Functional Interaction Network and external interaction databases such as IntAct, BioGRID, ChEMBL, iRefIndex, MINT and STRING. Our Pathway and Expression Analysis tools enable ID mapping, pathway assignment and overrepresentation analysis of user-supplied data sets. To support pathway annotation and analysis in other species, we continue to make orthology-based inferences of pathways in non-human species, applying Ensembl Compara to identify orthologs of curated human proteins in each of 20 other species.

The resulting inferred pathway sets can be browsed and analyzed with our Species Comparison tool. Collaborations are also underway to create manually curated data sets on the Reactome framework for chicken, *Drosophila* and rice.

INTRODUCTION

Reactome is an open source, open access, manually curated, peer-reviewed pathway database of human pathways and processes (1). Pathway annotations are created by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to proteins (UniProt) and genes (NCBI EntrezGene, Ensembl, UCSC and HapMap), small molecules (KEGG Compound and ChEBI), primary research literature (PubMed) and GO controlled vocabularies (2–9). The Reactome data model generalizes the concept of a reaction to include transformations of entities such as transport from one compartment to another and interaction to form a complex, as well as the chemical transformations of classical biochemistry. Entities include nucleic acids, small molecules, proteins (with or without post-translational modifications) and macromolecular complexes. This generalization permits the capture of a range of biological processes that spans signaling, metabolism, transcriptional regulation, apoptosis and synaptic transmission in a single internally consistent, computationally navigable format. Reactome is an

*To whom correspondence should be addressed. Tel: +1 416 673 8514; Fax: +1 416 977 1118; Email: lincoln.stein@gmail.com
Correspondence may also be addressed to Peter D'Eustachio. Tel: +1 212 263 5779; Fax: +1 212 263 8166; Email: Peter.D'Eustachio@nyumc.org

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

© The Author(s) 2010. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.5>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

all-inclusive resource of human pathways for basic research, genome analysis, pathway modeling, systems biology and education. In the past 2 years, the Reactome data set has nearly doubled in size and new tools for data aggregation and data analysis have become available. To support the continued development of the Reactome knowledgebase, we have redesigned the Reactome web site and data analysis software.

Expanded pathway coverage

Reactome's recruitment of expert authors and curators has given us access to key aspects of human biology. The current release of Reactome (Version 34, September 2010) describes the roles of 5272 human proteins (26% of the 20 286 human SwissProt entries) and 3504 macromolecular complexes in 3847 reactions organized into 1057 pathways. Over the last year, we have added new higher order topics. Notable additions include the molecular anatomy of transcriptional regulation, a largely complete catalogue of receptors with known ligands involved in GPCR signaling (10), Toll-like Receptors, Chromosome maintenance, Olfactory Signaling, Myogenesis, N-Glycan biosynthesis and Metabolism of RNA. Reactome has prototyped additional types of annotations to support pathway curation. We have curated pathways relating to insulin signaling cycle to prototype pathway-disease annotations. We have also developed a framework for physiological process annotations, e.g. vesicle transport and glutamate-mediated neurotransmission. To support the creation of the new pathway diagrams, we defined canonical pathways corresponding to 161 discrete biological domains. This enabled a simplification of the event hierarchy in the new Pathway Browser and has minimized event sharing across different pathways. Visualization features were implemented in the Author and Curator Tool to enable the layout and editing of these new pathways diagrams as part of our curation process.

The Reactome data model has been extended to the manual curation of pathways in model organisms. Gallus Reactome (<http://gallus.reactome.org>), an effort led by Carl Schmidt of the University of Delaware has been modeled after Reactome. The first public release was in mid-2009 and now Gallus Reactome includes annotations for 127 reactions involving 133 *Gallus* proteins in the domains of intermediary metabolism and DNA repair. A collaboration among Reactome, Michael Ashburner and Mark Williams at Cambridge University, similarly uses Reactome software and hardware to create and maintain a *Drosophila* pathway database (<http://fly.reactome.org/>). Its third release went public in mid-2010, and includes data for Wingless, JAK/STAT, Imd, Toll, Hedgehog, Circadian Clock, Hippo/Warts and Planar Cell Polarity signaling pathways.

Redesigned reactome web site

The rapid growth in content has made the 'starry sky' reaction map display unwieldy as a navigation and visualization tool. At the same time, our outreach has grown to encompass diverse user groups interested in browsing a particular process or protein as a textbook, analyzing

high-throughput expression data sets, data mining and data aggregation and online resources for education. Our front page has thus been redesigned to support quick, intuitive access to our data and tools as both features of the knowledgebase continue to grow. The new web site retains a comprehensive top menu bar that provides access to all of our tools and resources. It is now accompanied by a sidebar that provides access to basic, widely used tools for pathway browsing and data analysis, and panels that give a thumbnail overview of Reactome information, tutorials, recent news and a view of a recently added pathway of topical interest.

Improved pathway visualization

Visualization of full pathway information in a consistent format is vital to support the pathway-based analysis of complex experimental and computational data sets. To support such visual navigation and analysis of Reactome data we have developed, in collaboration with the ENFIN project (11), a new Pathway Browser based upon the Systems Biology Graphical Notation (SBGN) (12). SBGN is a standard graphical representation of biological pathway and network models, i.e. every molecule and reaction has a particular shape, color and cellular location. Our entire content has been organized into 161 canonical pathways, each displayed in this format. The Reactome Pathway Browser consists of four key elements. First, the 'Search' bar at the top of the page queries the entire Reactome database. Second, the 'Pathways' panel provides a scrolling display of the Reactome canonical pathway hierarchy. Third, clicking on the pathway name displays the corresponding pathway diagram in the 'Visualization' panel on the right side. The 'Visualization' panel offers interactive and dynamic pathway diagrams permitting zooming, scrolling and highlighting of events and molecules. Fourth, clicking on events and molecules in the pathway diagram uncovers a 'Details' panel below the pathway diagram with additional textual information about the events and molecules, respectively. Further functionality is provided in the form of context sensitive menus within the 'Visualization' panel (Figure 1). The precise features of the context sensitive menus are determined by the nature of the physical entity (small molecule, protein, complex): (i) a list of the other pathways in Reactome in which the selected entity participates; (ii) a display of the physical entities that contribute to the macromolecular complex; and, optionally (iii) a list of interactors of the entity from selected interaction databases (described later). The pathway diagrams are available for download as static PNG and PDF files. Dynamic pathway images compatible with third-party tools like Cytoscape (13) and CellDesigner (14) are currently being developed.

Integrating molecular interactions and networks onto reactome pathway diagrams

The Reactome data sets are a highly reliable platform for pathway-based data analysis but suffer from a low coverage of human proteins. To increase protein coverage and associated protein annotations, we have integrated

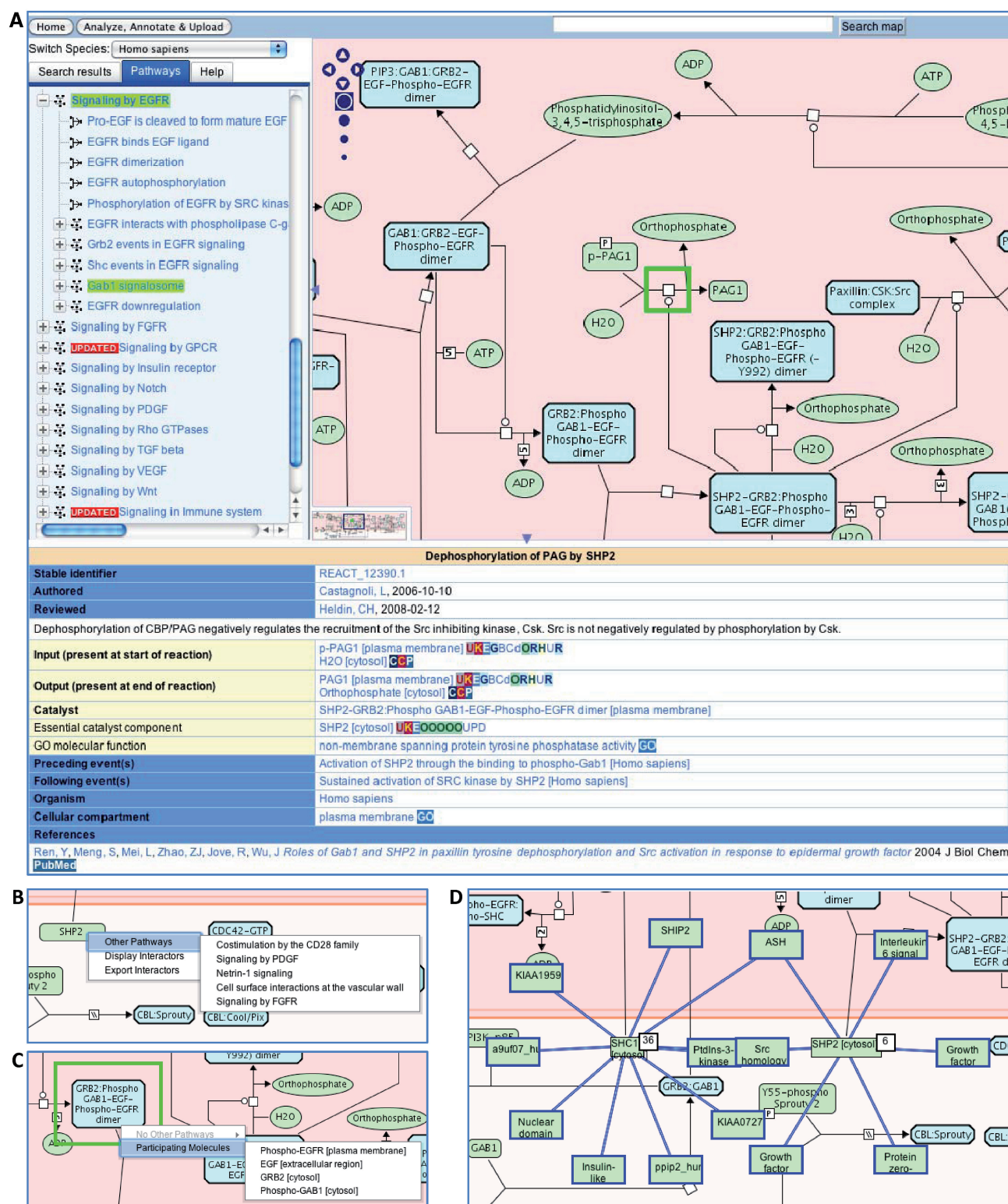


Figure 1. The Reactome pathway browser and the molecular overlays. (A) The main features of the pathway browser are the 'Search' bar at the top, the 'Pathways' panel on the left, the 'Visualization' panel on the right and a 'Details' panel (hidden by default when the pathway browser first loads). When a pathway is selected from the 'Pathways' panel hierarchy, it is highlighted in bright green in the hierarchy and its parent terms are highlighted in yellow-green. A description of any reaction or pathway in the pathway diagrams can be displayed below the diagram in the 'Details' panel by selecting the event within the diagram. Right-clicking the mouse while the cursor is over a physical entity displays the context sensitive menus. (B) Selecting 'Other Pathways' displays additional Reactome pathways that contain the highlighted physical entity. Navigating to the additional pathway is achieved by clicking the pathway name. (C) Selecting 'Participating Molecules' displays additional components of the macromolecular complex. Displaying the component information in the 'Details' panel is achieved by clicking the entity name. (D) Protein-protein interactions are displayed in the pathway browser for SHC1 [cytosol] and SHP2 [cytosol] physical entities.

molecular interaction (MI) data and network information into the Reactome pathway diagrams (Figure 1). The MI overlay displays proteins interacting with the manually annotated protein components of a Reactome pathway. As mentioned previously, individual protein interactors can be displayed using the context-dependent menus in the pathway 'Visualization' workspace. It is also possible to overlay all the interactors for all the pathway proteins by means of the 'Analyze, Annotate and Upload' feature of the Pathway Browser. The network overlay tool employs a PSICQUIC interface to implement flexible import of binary MI data into Reactome pathway diagrams. PSICQUIC is already widely implemented by interaction databases, including BioGRID, ChEMBL, IntAct, iRefIndex, MINT and STRING (15–20). The nodes and edges of the network overlay are interactive, providing links to the physical entity and interaction databases, respectively. Two of additional interaction data sets are managed by the Reactome group, 'Reactome' and 'Reactome-FIs'. The original 'Reactome' data set reflects MI data derived from Reactome reactions and complexes. A new 'Reactome-FIs' (functional interactions) data set unites interactions from Reactome and those derived from other pathway databases, including KEGG, BioCyc, Panther, The Cancer Cell Map (<http://cancer.cellmap.org/>) and PID (7,21–23) with pair-wise interactions gleaned from physical protein–protein interactions

in human and model organisms, gene co-expression data, protein domain–domain interactions, protein interactions generated from text mining and GO annotations (24). The 'Reactome-FIs' network contains 209 988 functional interactions encompassing 10 956 proteins (excluding splice isoforms), reflecting 46% of SwissProt proteins.

Improved orthology prediction and visualization of predicted pathways for model organisms

Comparative analysis of biological processes offers important information on their evolution, and supports metabolic engineering, the study of human disease and the identification of potential drug targets. Curated human reactions were used previously to electronically infer reactions by orthology in 20 evolutionary divergent species, with the assistance of the OrthoMCL (25). To align Reactome more closely with the Ensembl set of genome data and genome analysis tools, we have shifted to Ensembl Compara (26) to support orthology-based reaction inferences in 20 species for which high-quality whole-genome sequence data are available, including all 12 of the species in the GO Reference Genome annotation project (27). Viewing diagrams for predicted pathways in another species are available from within the Pathway Browser (Figure 2). A new Species Comparison tool allows users to compare these predicted pathways with

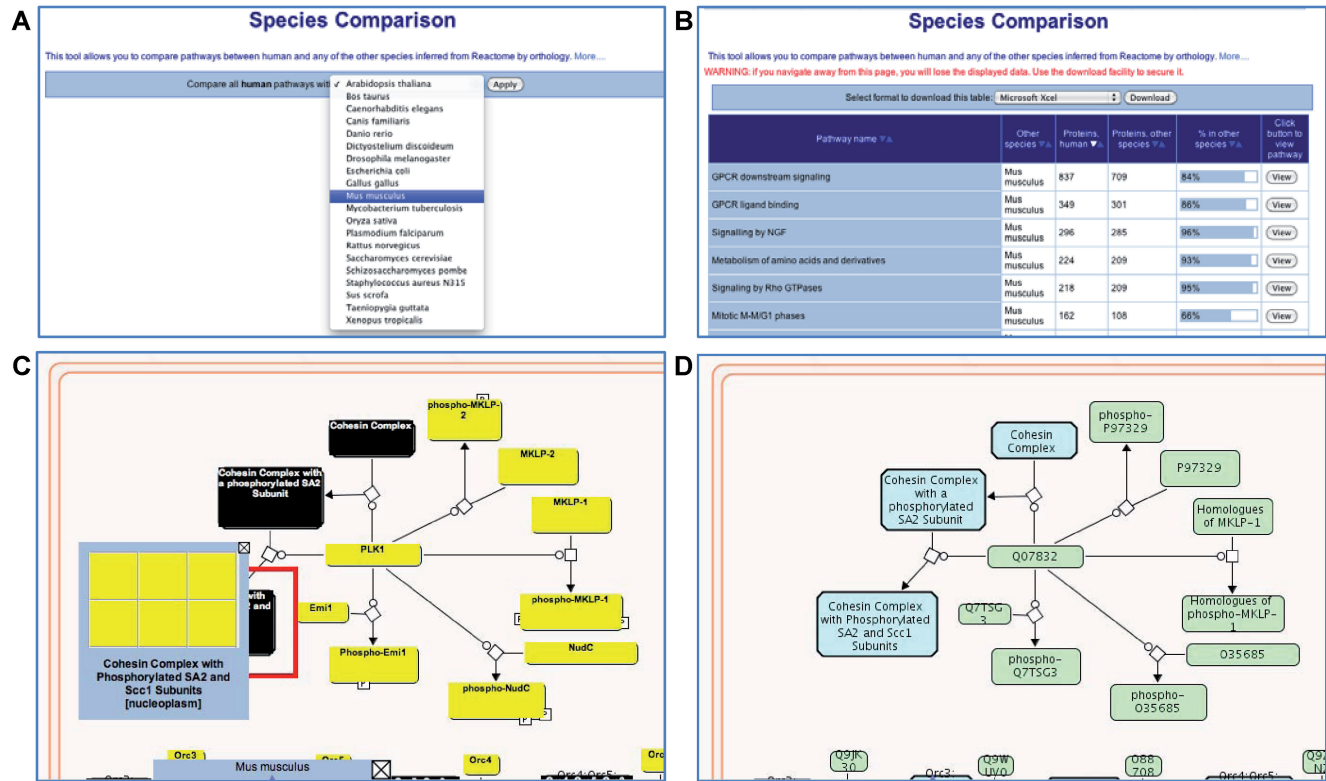
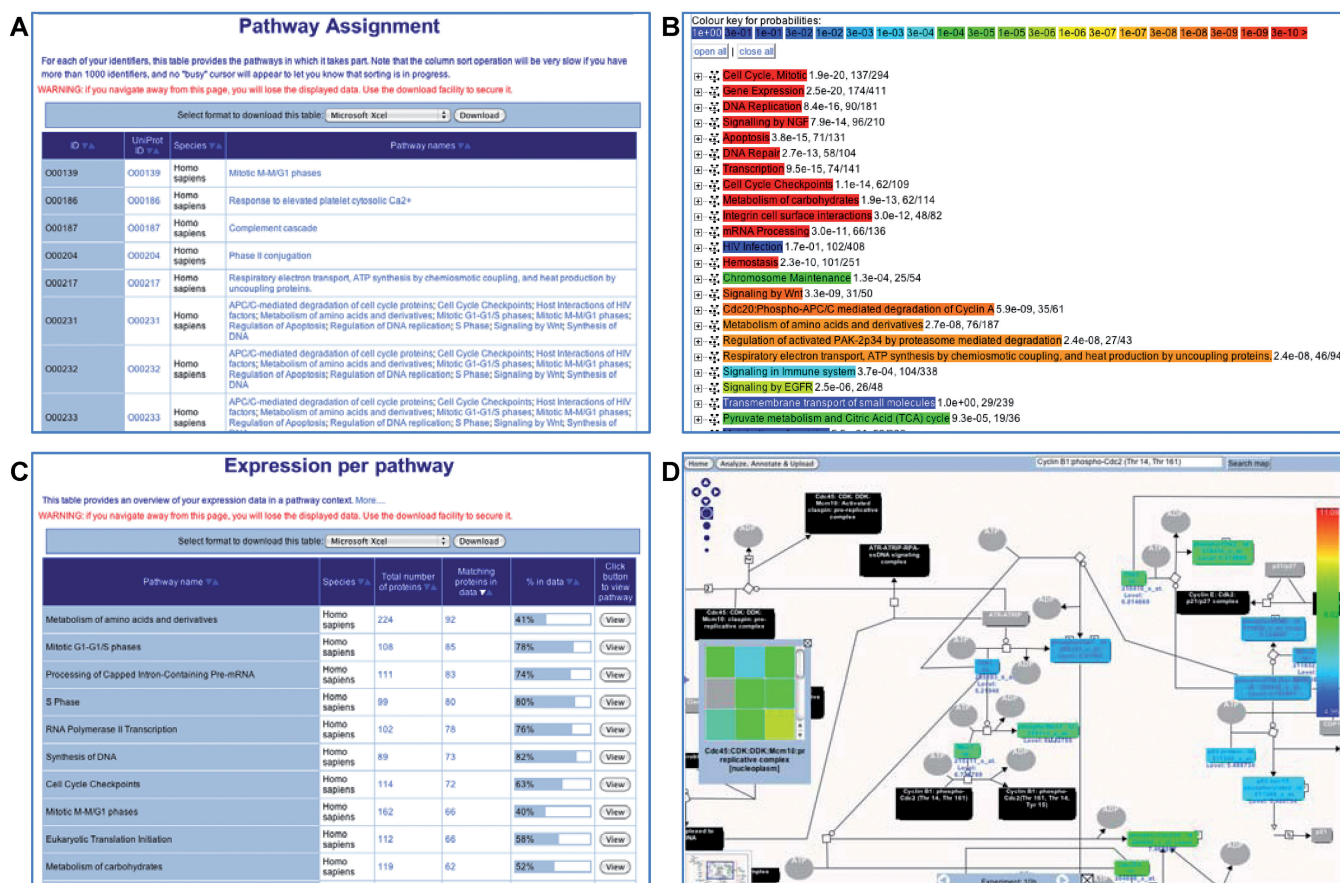


Figure 2. Species comparison tool and model organism pathway diagrams. (A) A drop-down menu is used to select the model organism species. (B) Results for the comparison of human and mouse pathways. Each row in the table is a pathway; the columns are pathway name, model organism name, number of proteins in the human pathway, number of orthologous proteins in the inferred model organism pathway, a graphical representation of the ratio of these two values and a 'View' button that creates a pathway diagram. 'Sort' buttons at the top of each column allow the table to be re-ordered according to cell contents in that column. (C) The pathway browser displaying the comparison of the human and mouse Mitotic M-M/G1 pathways. Physical entities in the pathway diagram are color-coded: gray, no match; black, a complex (multicomponent) entity; and yellow, the protein's ortholog is present in human. (D) The pathway browser displaying the inferred mouse Mitotic M-M/G1 pathway.



change or statistical values) and shows how expression/abundance levels affect reactions and pathways in living organisms (Figure 3). Again, the results are provided in a tabular format. Results from both the pathway and expression analysis results can be downloaded as a spreadsheet or tab- and comma-separated formats. The colored pathway diagrams can be downloaded in publication quality format. The molecular overlay and context sensitive menu features are also enabled in the colored pathway diagrams, providing links from user-supplied experimental data to Reactome pathways and MIs and networks.

Other collaborations and data exports

In collaboration with NCBI, Reactome annotations of pathways are being deposited into the NCBI BioSystems database, a large data repository for cataloguing molecules (nucleic acids, proteins, small molecules, drugs, etc.) that interact in biological systems (28). Reactome is part of the BioPAX Consortium to develop a data-exchange language to describe pathways, reactions and interactions (29). We have partnered with Gene Set Enrichment Analysis (GSEA) group at the Broad Institute to expand the collection of curated gene sets in the Molecular Signatures Database (MSigDB) to include Reactome's high-quality pathway data (30). Reactome participated in this year's Google Summer of Code program, collaborating with WikiPathways (31). The integration of pathway and interaction data has been a key element of the Reactome redevelopment. We have provided a new file format for the exchange of binary interaction data, based upon the PSI-MITAB format (32). In response to user requests, we recently changed the representation of protein modifications in Reactome to the PSI-MOD standard (33). We continue to support the use of Reactome data for ontology development with our relationships with the Gene and Protein Ontology groups. Reactome web pages link out to many online bioinformatics databases. This year, additional cross-references to RSCB Protein Data Bank (34), Comparative Toxicogenomics Database (35), DockBlaster (36), BioGPS (37) and dbSNP (38) have been added to the protein pages. Reactome software and data are now distributed under the terms of a Creative Commons Attribution 3.0 Unported License, that grants parties the non-exclusive right to use, distribute and create derivative works based on Reactome, provided that the software and information is correctly attributed to CSHL, OICR and EBI.

ACKNOWLEDGEMENTS

We are grateful to the many researchers who have volunteered to be external authors and reviewers. Development of the Reactome data model and fly and chicken databases is a collaborative project and this work benefited greatly from our interactions with Carl Schmidt, Mark Williams and Michael Ashburner.

FUNDING

National Human Genome Research Institute at the National Institutes of Health (grant number P41 HG003751); European Union 6th Framework Programme 'ENFIN' (grant number LSHG-CT-2005-518254). Funding for open access charge: Ontario Institute for Cancer Research.

Conflict of interest statement. None declared.

REFERENCES

- Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
- Ontology Consortium Gene. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
- Degtyarenko,K., Hastings,J., de Matos,P. and Ennis,M. (2009) ChEBI: an open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinform.*, Chapter 14: Unit 4.9.
- Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
- Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
- Jain,E., Bairoch,A., Duvaud,S., Phan,I., Redaschi,N., Suzek,B.E., Martin,M.J., McGarvey,P. and Gasteiger,E. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **10**, 136.
- Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
- Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
- Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.
- Jassal,B., Jupe,S., Caudy,M., Birney,E., Stein,L., Hermjakob,H. and D'Eustachio,P. (2010) The systematic annotation of the three main GPCR families in Reactome. *Database*, 2010, doi:10.1093/database/baq018.
- Kahlem,P., Clegg,A., Reisinger,F., Xenarios,I., Hermjakob,H., Orengo,C. and Birney,E. (2009) ENFIN—A European network for integrative systems biology. *C. Roy. Biol.*, **332**, 1050–1058.
- Le Novère,N., Hucka,M., Mi,H., Moodie,S., Schreiber,F., Sorokin,A., Demir,E., Wegner,K., Aladjem,M.I., Wimalaratne,S.M. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.
- Killcoyne,S., Carter,G.W., Smith,J. and Boyle,J. (2009) Cytoscape: a community-based framework for network modeling. *Methods Mol. Biol.*, **563**, 219–239.
- Funahashi,A., Jouraku,A., Matsuoka,Y. and Kitano,H. (2007) Integration of CellDesigner and SABIO-RK. *In Silico Biol.*, **7**, S81–S90.
- Aranda,B., Achuthan,P., Alam-Farouque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
- Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. *et al.* (2008) The BioGRID interaction database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.
- Ceol,A., Chatr Aryamontri,A., Licata,L., Peluso,D., Briganti,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2009) MINT, the

- molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
18. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. *et al.* (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
 19. Overington, J. (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput. Aided Mol. Des.*, **23**, 195–198.
 20. Razick, S., Magklaras, G. and Donaldson, I.M. (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, **9**, 405.
 21. Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F., Kaipa, P., Karthikeyan, A.S., Kothari, A., Krummenacker, M. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
 22. Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, **38**, D204–D210.
 23. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
 24. Wu, G., Feng, X. and Stein, L. (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.*, **11**, R53.
 25. Chen, F., Mackey, A.J., Stoeckert, C.J. Jr and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
 26. Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
 27. Reference Genome Group of the Gene Ontology Consortium. (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol.*, **5**, e1000431.
 28. Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W. and Bryant, S.H. (2010) The NCBI BioSystems database. *Nucleic Acids Res.*, **38**, D492–D496.
 29. Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.
 30. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. and Mesirov, J.P. (2007) GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics*, **23**, 3251–3253.
 31. Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R. and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
 32. Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A.F., Vinod, N., Bader, G.D., Xenarios, I., Wojcik, J., Sherman, D. *et al.* (2007) Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.
 33. Montecchi-Palazzi, L., Beavis, R., Binz, P.A., Chalkley, R.J., Cottrell, J., Creasy, D., Shofstahl, J., Seymour, S.L. and Garavelli, J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.
 34. Dutta, S., Burkhardt, K., Young, J., Swaminathan, G.J., Matsuura, T., Henrick, K., Nakamura, H. and Berman, H.M. (2009) Data deposition and annotation at the worldwide protein data bank. *Mol. Biotechnol.*, **42**, 1–13.
 35. Wieggers, T.C., Davis, A.P., Cohen, K.B., Hirschman, L. and Mattingly, C.J. (2009) Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC Bioinformatics*, **10**, 326.
 36. Irwin, J.J., Shoichet, B.K., Mysinger, M.M., Huang, N., Colizzi, F., Wassam, P. and Cao, Y. (2009) Automated docking screens: a feasibility study. *J. Med. Chem.*, **52**, 5712–5720.
 37. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W. 3rd *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
 38. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.