20.453J / 2.771J / HST.958J Biomedical Information Technology
Fall 2008

PID: The Pathway Interaction Database

Carl F. Schaefer[1], Kira Anthony[2], Shiva Krupa[3], Jeffrey Buchoff[4], Matthew Day[2], Timo Hannay[2], Kenneth H. Buetow[1]

[1]National Cancer Institute, Center for Biomedical Informatics and Information Technology

[2]Nature Publishing Group

[3]Novartis Institutes for Biomedical Research, Inc., Novartis Knowledge Center

[4]SRA International, Inc., Health Research and Informatics

KEYWORDS: pathway, database, cell signaling, signal transduction

ABSTRACT

The Pathway Interaction Database (PID, http://pid.nci.nih.gov) is a freely available collection of curated and peer-reviewed pathways composed of human molecular signaling and regulatory events and key cellular processes. Created in a collaboration between the U.S. National Cancer Institute and Nature Publishing Group, the database serves as a research tool for the cancer research community and others interested in cellular pathways, such as neuroscientists, developmental biologists, and immunologists. PID offers a range of search features to facilitate pathway exploration. Users can browse the predefined set of pathways or create interaction network maps centered on a single molecule or cellular process of interest. In addition, the batch query tool allows users to upload long list(s) of molecules, such as those derived from microarray experiments, and either overlay these molecules onto predefined pathways or visualize the complete molecular connectivity map. Users can also download molecule lists, citation lists and complete database content in extensible markup language (XML) and Biological Pathways Exchange (BioPAX) Level 2 format. The database is updated with new pathway content every month and supplemented by specially commissioned articles on the practical uses of other relevant online tools.

INTRODUCTION

The Pathway Interaction Database (PID, http://pid.nci.nih.gov) is a growing collection of human signaling and regulatory pathways curated from peer-reviewed literature and stored in a computable format. PID was designed to deal with two issues affecting the representation of biological processes: the arbitrariness of pathway boundaries and the need to capture knowledge at different levels of detail. Pathway boundaries are often arbitrary and overlapping: different biologists might include different biochemical interactions in, for example, “the p53 signaling pathway”; and it is not unusual for two pathways representing distinct processes to have one or more interactions in common. This fuzziness simply reflects the fact that terms like “the p53 signaling pathway” and “the BCR signaling pathway” are high-level concepts of convenience, designating slices through the very complex mix of concurrent processes in the cell. An important goal of PID is to provide an operational definition of high-level concepts like “the BCR signaling pathway” (Figure 1) in the form of predefined pathways, while at the same time allowing a user to explore novel networks composed computationally from the universe of interactions underlying the predefined pathways. Current knowledge of the components of any given biological process is uneven. For example, for some protein interactants the precise post-translational modifications might be known, while for other interactants perhaps the only sure knowledge is that the protein is “active”. PID provides descriptive mechanisms to cover both of these cases. The ability to represent information at different levels of detail is also useful in communicating generalizations. For example, it is sometimes useful to encapsulate a complex process such as cytoskeleton reorganization as a single event or to treat as a single entity a set of proteins such as Class 1A PI3Ks that are functionally equivalent in catalyzing a given event. PID has mechanisms for dealing with incomplete knowledge, for encapsulating complex events and for expressing generalizations.

PID has adopted a network-level representation, similar to Reactome (1), HumanCyc (2), and KEGG (3). Like Reactome and HumanCyc, PID annotates interactions with citations to the literature. PID differs from Reactome, HumanCyc, and KEGG in its focus on signaling and regulatory pathways; it does not attempt to cover metabolic processes or generic mechanisms like transcription and translation. PID contains only structured data and it links to but does not reproduce molecular information readily available from other sources, such as nucleotide or amino acid sequence, molecular weight, and chemical formula. The principal source of data in PID is the highly curated "NCI-Nature Curated" collection of pathways, but PID also includes two other sources of data: data imported into the PID data model from Reactome's Biological Pathways Exchange (BioPAX) Level 2 (4) export, and an import of information from the BioCarta collection of pathways (Table 1). All data in PID is freely available, without restriction on use. Bulk downloads are available in BioPAX Level 2, a standard format for exchanging pathway information, and a PID-specific XML format at http://pid.nci.nih.gov/PID/download.shtml.

DATA MODEL

In PID, an interaction is an event with its participating molecules and conditions. A PID pathway is a network of these events connected by the participant molecules. PID recognizes four kinds of molecules: small molecules (called compounds), RNA, proteins, and complexes. PID recognizes five kinds of events: gene regulation (called transcription, but encompassing both transcription and translation), molecule transport (called translocation), small-molecule conversion (called reaction), protein-protein interactions (called modification), and black-box processes whose internal composition is not provided (called macroprocesses). In addition, an entire pathway can be abstracted and used as a single event in another pathway. As a participant in an event, a molecule may have one of four roles: input, output, positive regulator, and negative regulator. These roles define simple relations: an interaction consumes its inputs (but not its regulators) and produces its outputs; and the inputs, positive regulators and the absence of negative regulators are jointly the necessary and sufficient causes of the interaction.

Each molecule in PID has a defining entity, called a basic molecule. Basic molecules are distinguished by their nucleotide or amino acid sequence (for macromolecules) or by their chemical formula (for small molecules). While PID does not record the sequence of a macromolecule or the chemical formula for a small molecule, each protein or RNA is associated with a UniProt or Entrez Gene accession and most small molecules are associated with Chemical Abstracts Service (CAS) registry numbers. A basic molecule has a primary name and may have multiple aliases. Each molecule use, as an interactant in an interaction or as a component of a complex, references its corresponding basic molecule. Each molecule use may have additional information, including post-translational modifications (for proteins) and cellular location and activity state (for all molecule types).

A basic protein molecule has a single identifying UniProt accession associated with a particular amino acid sequence. If the particular isoform of a protein used in an interaction is not known, then the basic protein molecule may be associated with an Entrez Gene identifier instead of a UniProt accession; in PID, this method of identifying proteins is restricted almost entirely to the

uncurated section of the database imported from BioCarta. A use of a protein as a participant in an interaction or component of a complex may have additional attributes: post-translational modifications, an abstract activity-state attribute, and a cellular location attribute. Currently, PID uses 13 types of post-translational modifications, with phosphorylation being by far the most frequently used modification (Table 2). The abstract activity-state attribute, with values such as “active” and “inactive”, allows curators to distinguish functionally different forms of a protein even if the precise covalent modifications are not known. Values for the cellular location attribute are drawn from the Gene Ontology (GO) Cellular Component vocabulary (5). Cleaved subunits of a precursor protein are not distinguished by the post-translational modification mechanism; rather they are treated as basic protein molecules separate from each other and from the precursor. However, PID explicitly relates the cleaved subunit to its precursor and records the cleavage coordinates when these are known. A PID protein corresponds roughly to a BioPAX Level 3 proteinReference, while a BioPAX Level 3 protein corresponds to a PID protein use (with post-translational modifications and cellular location).

PID allows the definition of generic proteins, complexes, small molecules, and RNA molecules. A generic molecule is called a family, but is not restricted to the traditional protein families defined by sequence similarity: any set of proteins (or other type of molecule) that are in some respect functionally equivalent may be grouped in a family. Individual protein members of a protein family may have post-translational modifications or activity-states. The family itself can be used as a participant in an interaction, or as a component of a complex.

Because data are entered by multiple curators and because the database contains data from multiple sources, PID needs rules for determining equivalence of molecules. Two basic molecules that are neither families nor complexes are equivalent if they have the same external database accession (e.g., UniProt or Entrez Gene), or if, in cases where neither has an external database accession, they have the same name. Two molecule uses (as participant in an interaction or component of a complex or member of a family) are equivalent if they refer to the same basic molecule, and have the same (or no) post-translational modifications, and have the same (or no) activity-state attribute, and have the same (or no) cellular location attribute. Two basic families (or complexes) are equivalent if they have the same number of members (or components) and if for each member (component) of one, there is an equivalent member (component) in the other. These rules are applied recursively to define, for example, equivalent uses of complexes with components that are families. Equivalence of molecule uses is the basis on which novel networks are constructed: any two interactions in the database may be joined in a network if one interaction has a participant that is equivalent to a participant in the other interaction. Analogous rules of equivalence are implemented for interactions and entire networks, allowing equivalent (redundant) interactions to be pruned from the novel networks.

An interaction may be supported by one or more citations to the literature. Currently, 3233 of the 4293 interactions in the NCI-Nature Curated data source are annotated with at least one of 2957 unique PubMed references. In addition, an interaction may be annotated with one or more evidence codes that specify the kind of evidence adduced in the citations in support of the interaction (Table 3).

A predefined pathway is a curated pathway representing a known biological process. At present, every pathway stored in the PID database is a predefined pathway and every interaction in the database is a member of at least one predefined pathway. However, the search and retrieval tools allow the user to compose novel pathways from interactions defined in the predefined pathways. This ability to recombine interactions and to thus create novel pathways is a distinguishing feature of PID.

DATA CURATION

Nature Publishing Group (NPG) editors create the NCI-Nature Curated pathways. Pathways selected for curation are based on potential drugs targets, suggestions made by users and reviewers, and other molecules known to be of interest to the cell signaling community. A list of NCI-Nature Curated pathways, along with a list of the pathways imported from Reactome and BioCarta, can be found on the Browse pathways page of the PID website: http://pid.nci.nih.gov/PID/browse_pathways.shtml

In curating, editors synthesize meaningful networks of events into defined pathways and adhere to the PID data model for consistency in data representation: molecules and biological processes are annotated with standardized names and unambiguous identifiers; and signaling and regulatory events are annotated with evidence codes and references. To ensure accurate data representation, editors assemble pathways from data that is principally derived from primary research publications. The majority of data in PID is human; however, if a finding discovered in another mammal is also deemed to occur in humans, editors may decide to include this finding, but will also record that the evidence was inferred from another species. Prior to publication, all pathways are reviewed by one or more experts in a field for accuracy and completeness.

WEB INTERFACE AND APPLICATION

PID provides several query options: a simple query, an advanced query, a connected molecules query, and a batch query. In the simple query, the user provides the name, alias, or accession of a molecule or biological process; wildcarding is permitted. The query will return a list of all uses of the molecule, as simplex or as participant in a complex, and all uses of the biological process, in the database, with hyperlinks to visualizations of the relevant predefined pathways containing the queried entities. The user also has the option to visualize the novel network(s) that include all interactions using the queried entities. The advanced query allows the user to construct the set of novel networks from interactions that (a) involve any of a set of user-specified molecules, or (b) are part of any predefined pathway whose name includes a user-specified key word, or (c) have a user-specified GO Biological Process term or National Cancer Institute (NCI) Thesaurus (6) term as their event type or condition. An important feature of the advanced query is the provision for including interactions that are immediately upstream or downstream of the set of interactions retrieved by molecule, pathway name, or GO/NCI Thesaurus term. The connected molecules query allows a user to find a novel network that connects two or more molecules specified by name, alias, or accession. The query will find only one of the possibly many networks satisfying the constraint, but the one found will have the

minimum number of interactions. Finally, the batch query allows a user to upload one or two lists of molecule identifiers (name, alias, or accession). The user has two options: to analyze the number of molecules in the lists that "hit" each predefined pathway or to construct the novel network(s) that include all interactions using any of the listed molecules. For the first option, the query uses a hypergeometric distribution to compute the probability that each pathway in the database is hit by molecules in either of the lists. The query returns a list of pathways ordered by P-value. In the visualization of a predefined pathway (first option) or novel pathway (second option), molecules from the first list are colored blue, molecules from the second list are colored red, and any molecules appearing in both lists are colored purple. Supplementary Figure 1 presents an example of invoking the batch query with a single molecule list, the 120 protein kinases found by Greenman et al. (7) to have at least one cancer-predisposing mutation. Selecting the predefined pathways option, one can see that this list samples a small number of pathways, biased toward immune cell signaling, at a P-value < 0.0001.

While PID associates a single external database accession (typically a UniProt accession) with a protein, the query interface searches PID not only by UniProt accession, but also by related gene identifiers (HUGO symbol, alias, Entrez Gene identifier). Any predefined pathway or novel network can be visualized in either GIF (Graphics Interchange Format) or SVG (Scalable Vector Graphics) graphic mode. Network graphics are all automatically constructed from the underlying data using the GraphViz package (8). Events and molecule uses in the graphics are hyperlinked to HTML pages of information about the interaction or molecule. In addition, any predefined pathway or novel network can be exported in native PID XML or BioPAX Level 2 formats. Using the BioPAX export, a user can also visualize PID pathways in Cytoscape (http://cytoscape.org) a popular third-party network visualization tool (9). For any predefined pathway, the user can obtain (and export to tab-separated format) a list of literature citations and participating molecules.

DISCUSSION AND FUTURE DIRECTIONS

PID is a highly structured, curated database of molecular interactions and events that compose human cell signaling and regulatory pathways. A particular strength of PID is the ability to create novel networks that can reveal parallel alternative paths to events of interest, like activation of a protein or disassembly of a complex in the DNA repair process. In cancer biology, such a view can elucidate the variety of strategies that a given type of cancer may adopt, explain why a single-agent therapy is not effective, and suggest potential multi-agent therapies. Increasingly, molecular networks are recognized as frameworks for integrating and interpreting experimental data. For example, by using pathways as the integrating framework, The Cancer Genome Atlas project has mapped genomic abnormalities of different types -- copy number, mutation, and methylation – to a set of oncogenic processes (publication forthcoming). At present, most attempts to profile tumor subtypes have relied on DNA and RNA assays. However, as high-throughput proteomic methods improve, the kind of detailed information on post-translational modifications of proteins available in PID will be essential in mapping more accurately the state of a cell.

Consistent with its focus on interactions and events derived from curated signaling cascades and regulatory processes, PID does not at present include interaction data deriving from high-throughput protein-protein interaction experiments. This reflects not a judgment on the quality of high-throughput data but a recognition that there are databases specifically designed to provide access to this data (10, 11, 12). However, while it does not lead directly to the construction of signaling cascades, information from high-throughput protein-protein interaction experiments can be useful in interpreting the curated pathways and assessing their completeness. For example, a high-throughput protein-protein interaction experiment can identify an unexpected binding partner for a catalyst, suggesting the possibility that the in vivo presence of the partner can sequester the catalyst and thus turn off downstream interactions. In the future PID will allow users to take advantage of high-throughput protein-protein interaction data, either by allowing users to upload interaction sets to be added to the novel networks created by PID queries or by querying other data sources (such as Pathway Commons, http://pathwaycommons.org) as needed to support a user query. The PID data model is currently being integrated with NCI's Cancer Bioinformatics Infrastructure Objects model (caBIO) (13), thereby making PID data accessible on NCI's caGrid (14).

ACKNOWLEDGEMENTS

Funding for this work was provided by the National Cancer Institute.

REFERENCES

1. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., et al. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, 8:R39doi:10.1186/gb-2007-8-3-r39.

2. H1. P. Romero, J. Wagg, M.L. Green, D. Kaiser, M. Krummenacker, and P.D. Karp. (2004) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6:R2 R2.1-17.

3. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, 36, D480-D484.

4. Bader, G.D., Cary, M., Sander, C. (2006) BioPAX - Biological Pathway Data Exchange Format. *Encyclopedia of Genomics, Proteomics and Bioinformatics*, John Wiley & Sons, Ltd. DOI: 10.1002/047001153X.g408117.

5. The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nature Genet*., 25:25-29.

6. Fragoso G., de Coronado S., Haber M., Hartel F., Wright L. (2004) Overview and utilization of the NCI Thesaurus. *Comp. Funct. Genomics*, 5(8):648-54.

7. Greenman C., et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, 446;153-8.

8. Gansner, E.R. and North, S.C. (1999) An open graph visualization system and its applications. *Software – Practice and Experience*, 00(S1), 1-5.

9. Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*., 2(10):2366-82.

10. Chatr-aryamontri, A., Ceol, A., Montecchi-Palazzi, L., Nardelli, G., Schneider, M.V., Luisa Castagnoli, L., and Cesareni, G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.,* 35(Database issue):D572-D574; doi:10.1093/nar/gkl950.

11. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., 32:D452-455.

12. Breitkreutz, B.J., Stark C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bähler, J., Wood, V., Dolinski, K., Tyers, M. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res, .*Jan;36(Database issue):D637-40.

13. Covitz, P.A., Hartel, F., Schaefer, C., De Coronado, S., Fragoso, G., Sahni, H., Gustafson, S., Buetow, K.H. (2003) caCORE: a common infrastructure for cancer informatics. *Bioinformatics*, Dec 12;19(18):2404-12.

14. Oster, S., Langella, S., Hastings, S., Ervin, D., Madduri, R., Phillips, J., Kurc, T., Siebenlist, F., Covitz, P., Shanbhag, K., Foster, I., Saltz, J. (2008) caGrid 1.0: an enterprise Grid infrastructure for biomedical research. *J. Am. Med. Inform. Assoc.*, Mar-Apr;15(2):138-49.

FIGURE

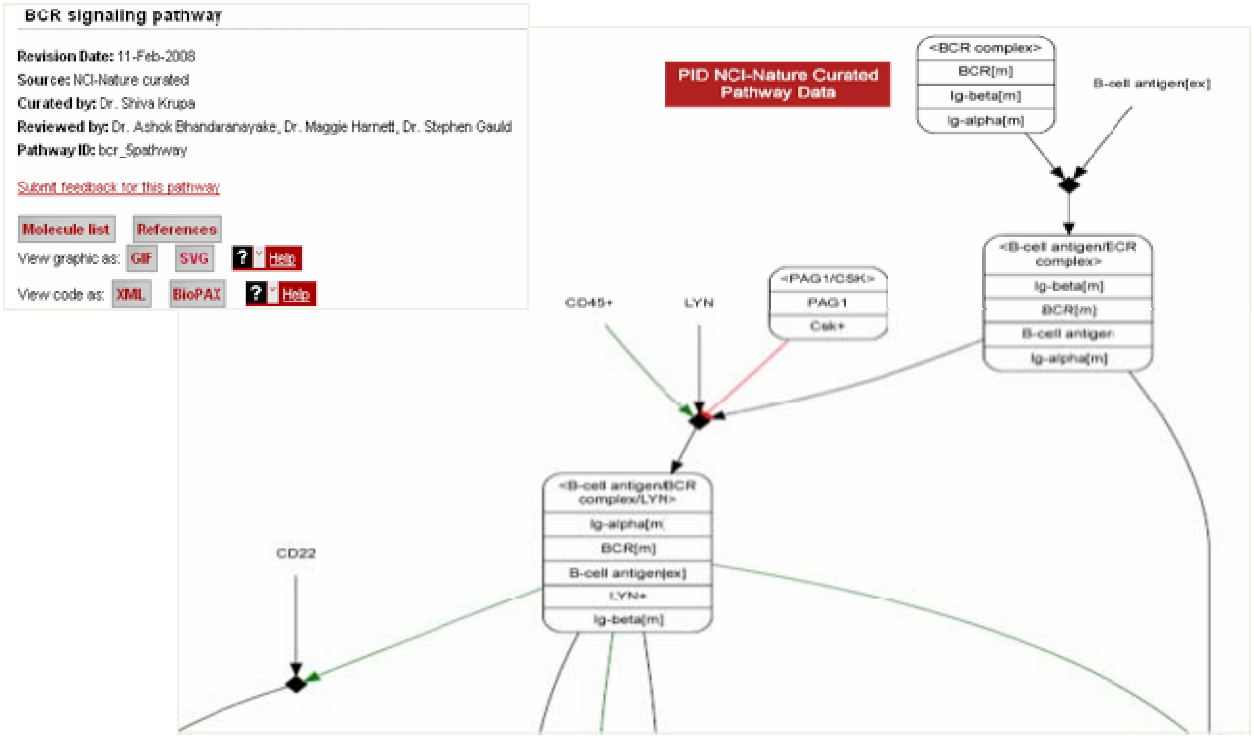


Figure 1. BCR signaling pathway: The pathway header information includes the date of the latest revision; the data curation or import source; the curator; the reviewers; the stable pathway identifier; links to a pathway-specific molecule list and a pathway-specific references list; and links to pathway graphic and text data exchange format options.

TABLES

	NCI-Nature Curated	Reactome Imported	BioCarta Imported
pathways	76	63	254
subpathways	40	753	0
interactions	4293	3466	3003
Proteins	2480	2692	4218
small molecules	129	617	205
complexes	1924	1897	880

Table 1: Summary of all data sources

Modification Type	All Uses	Unique Modifications
acetylation	111	46
farnesylation	7	4
geranylgeranylation	2	2
glycosaminoglycan	9	2
glycosylation	135	14
hydroxylation	13	3
methylation	13	2
myristoylation	15	4
oxidation	8	4
palmitoylation	52	14
phosphorylation	7251	1039
sumoylation	50	10
ubiquitination	146	52

Table2: Post-translational modifications in NCI-Nature Curated data source

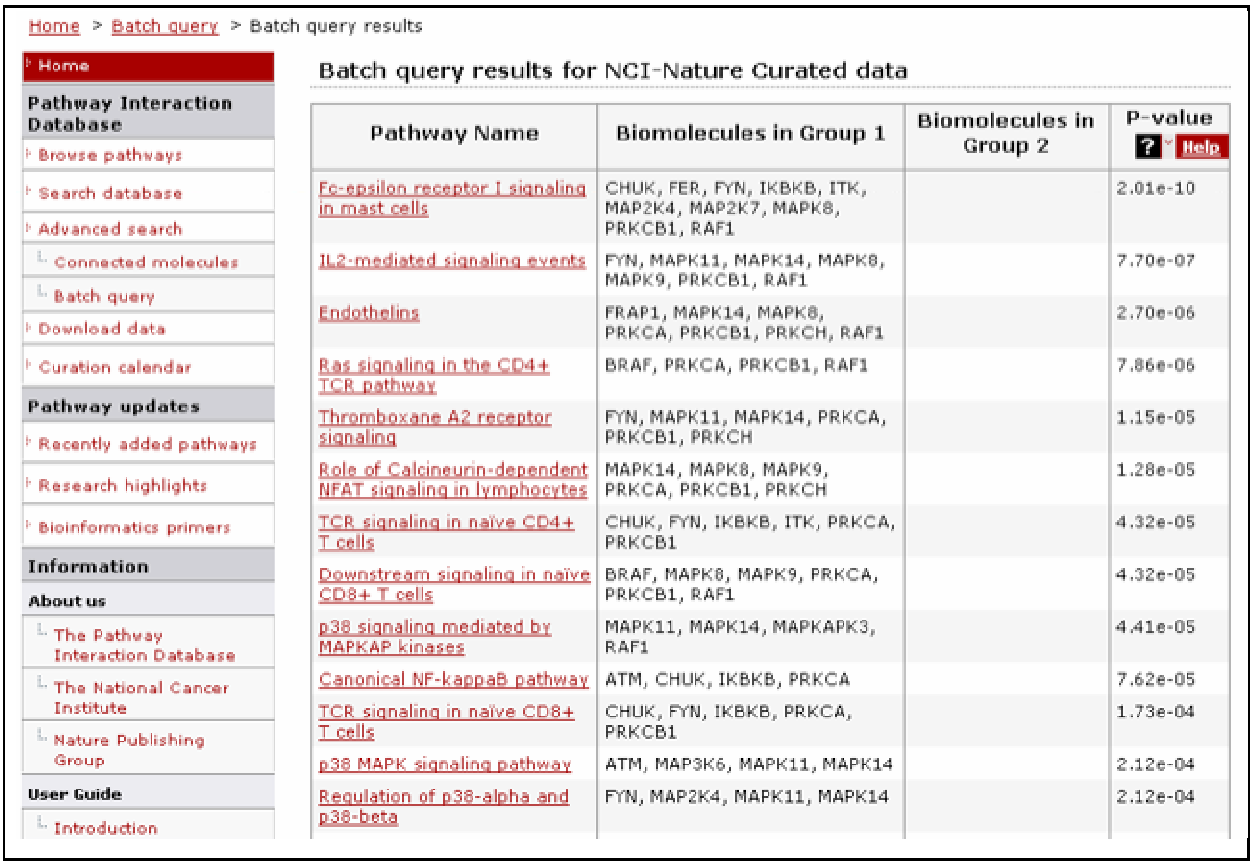
Code	Evidence Kind	Uses
IAE	Inferred from Array Experiments	2
IDA	Inferred from Direct Assay	1694
IEP	Inferred from Expression Pattern	27
IFC	Inferred from Functional Complementation	11
IGI	Inferred from Genetic Interaction	5
IMP	Inferred from Mutant Phenotype	1598
IOS	Inferred from Other Species	1077
IPI	Inferred from Physical Interaction	1144
RGE	Inferred from Reporter Gene Expression	278

Table 3: Evidence in NCI-Nature Curated data source

SUPPLEMENTARY DATA

SUPPLEMENTARY FIGURE LEGEND

Supplementary Figure 1. Batch query: The Batch query allows users to upload long list(s) of molecules identifiers (name, alias, or accession) and analyze the distribution of the molecules within predefined pathways (first option) or construct the complete set of network maps from interactions using any of the molecules (second option). Two lists can be uploaded simultaneously in order to compare data sets (e.g. genomic copy number alterations and somatic mutations). For the first option, the query uses a hypergeometric distribution to compute the probability that each pathway in the database is hit by molecules in either of the lists. The query returns a list of pathways ordered by P-value. In the example in the figure a list of 120 protein kinases containing at least one cancer-predisposing mutation (Greenman C., et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature, 446;153-8) was overlaid on the predefined pathways. The table below shows that this list of kinases samples 10 predefined pathways at a probability of $P < 0.0001$. A number of these pathways involve immune cell signaling.



Batch query results for NCI-Nature Curated data

Pathway Name	Biomolecules in Group 1	Biomolecules in Group 2	P-value ? Help
Fc-epsilon receptor I signaling in mast cells	CHUK, FER, FYN, IKBKB, ITK, MAP2K4, MAP2K7, MAPK8, PRKCB1, RAF1		2.01e-10
IL2-mediated signaling events	FYN, MAPK11, MAPK14, MAPK8, MAPK9, PRKCB1, RAF1		7.70e-07
Endothelins	FRAP1, MAPK14, MAPK8, PRKCA, PRKCB1, PRKCH, RAF1		2.70e-06
Ras signaling in the CD4+ TCR pathway	BRAF, PRKCA, PRKCB1, RAF1		7.86e-06
Thromboxane A2 receptor signaling	FYN, MAPK11, MAPK14, PRKCA, PRKCB1, PRKCH		1.15e-05
Role of Calcineurin-dependent NFAT signaling in lymphocytes	MAPK14, MAPK8, MAPK9, PRKCA, PRKCB1, PRKCH		1.28e-05
TCR signaling in naïve CD4+ T cells	CHUK, FYN, IKBKB, ITK, PRKCA, PRKCB1		4.32e-05
Downstream signaling in naïve CD8+ T cells	BRAF, MAPK8, MAPK9, PRKCA, PRKCB1, RAF1		4.32e-05
p38 signaling mediated by MAPKAP kinases	MAPK11, MAPK14, MAPKAPK3, RAF1		4.41e-05
Canonical NF-kappaB pathway	ATM, CHUK, IKBKB, PRKCA		7.62e-05
TCR signaling in naïve CD8+ T cells	CHUK, FYN, IKBKB, PRKCA, PRKCB1		1.73e-04
p38 MAPK signaling pathway	ATM, MAP3K6, MAPK11, MAPK14		2.12e-04
Regulation of p38-alpha and p38-beta	FYN, MAP2K4, MAPK11, MAPK14		2.12e-04