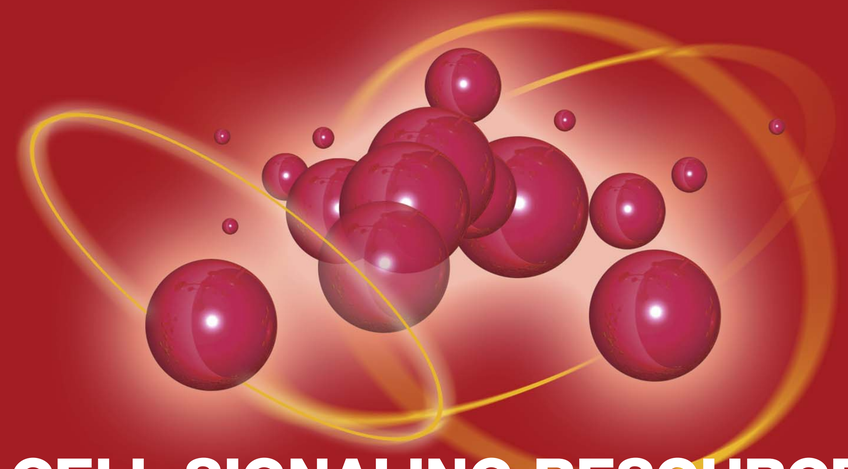
nature PathwayInteractionDatabase

Shiva Krupa[1], **Kira Anthony**[1], Jeffrey Buchoff[2], Matthew Day[3], Timo Hannay[4], Carl Schaefer[2]

1. Web Publishing, Nature Publishing Group, 25 First Street, Suite 104, Cambridge, MA, 02141, USA
2. National Cancer Institute Center for Bioinformatics and Information Technology, 2115 E Jefferson Street, Rockville, MD, 20852, USA
3. Web Publishing, Nature Publishing Group, The Macmillan Building, 4 Crinan Street, London, N1 9XW, UK
4. Nature.com, Nature Publishing Group, The Macmillan Building, 4 Crinan Street, London, N1 9XW, UK



A CELL SIGNALING RESOURCE

Abstract
The Pathway Interaction Database (**PID**, **http://pid.nci.nih.gov**) is a freely available collection of curated and peer-reviewed signaling pathways composed of human biomolecular interactions and cellular processes. Created in a collaboration between the U.S. National Cancer Institute and Nature Publishing Group, the database is a research tool for cell biologists, biochemists, computational biologists and bioinformaticians. The PID offers a range of tools to facilitate pathway exploration. Users can browse the pre-defined set of pathways and also create interaction network maps centered on a single molecule of interest or an extensive list of molecules. In addition, users can download complete data sets in extensible markup language (XML) and Biological Pathway Exchange (BioPAX) Level 2 formats. The database is updated every month and supplemented by a concise editorial section that provides synopses of recent noteworthy papers in cell signaling and specially commissioned articles on the practical uses of other relevant online tools. Users can sign up for free email alerts or RSS feeds to receive database updates.

Curation principles
- **Human model system**: We focus on human data. Interactions in other mammals that are inferred to occur in humans may be included with appropriate evidence codes.
- **Biological relevance**: Meaningful networks of undisputed interactions are synthesized into pathways. Pathways selected for curation are based on suggestions made by our users, potential drugs targets and other biomolecules we know to be of interest to researchers.
- **Authority**: Molecular interactions are identified in primary peer-reviewed literature. Editors judge whether an interaction is physiologically relevant and assign evidence codes to each interaction. All pathways are reviewed by experts in the field for accuracy and completeness.
- **Standardized nomenclature**: We use **HUGO** (**http://www.hugo-international.org/committee_nomen.htm**) gene symbols, **Entrez Gene** (**http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene**) aliases or **UniProt** (**http://www.pir.uniprot.org**) names or aliases for molecules. We use **Gene Ontology** (**GO, http://www.geneontology.org**) and **NCI Thesaurus** (**http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do**) controlled vocabulary terms as gene and gene product labels for molecules and pathway processes.

Data representation
A pathway is a set of interactions modeled as a directed graph with labeled nodes and edges.

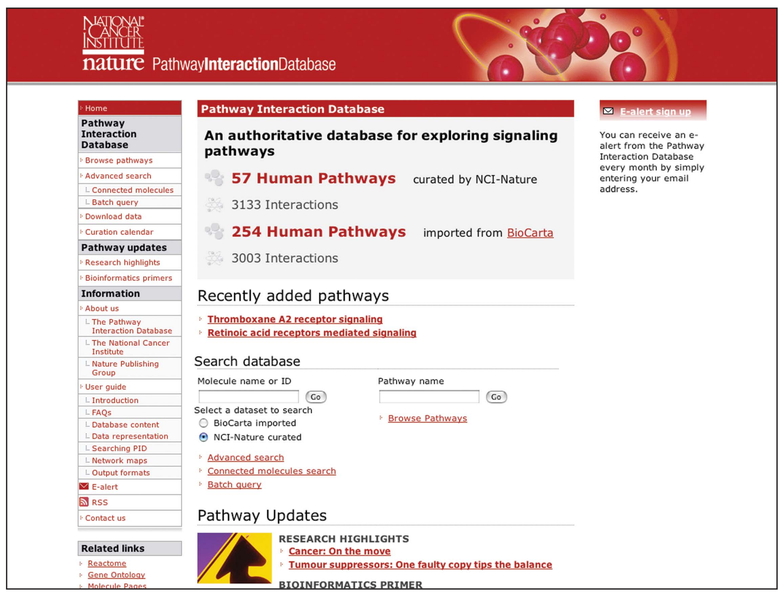
Pathway building blocks
Biomolecules (nodes) in PID may be **proteins**, **mRNA**, **small molecules** and **complexes**. Proteins are annotated with UniProt and Entrez Gene identifiers; mRNA is annotated with the Entrez Gene identifier; small molecules are annotated with CAS Registry Numbers; and complexes may be annotated with the GO cellular component identifier for the complex, if available.
Additional biomolecular nomenclature may include **activity states**, **post-translational modifications**, and GO terms for **location** and **molecular function**.

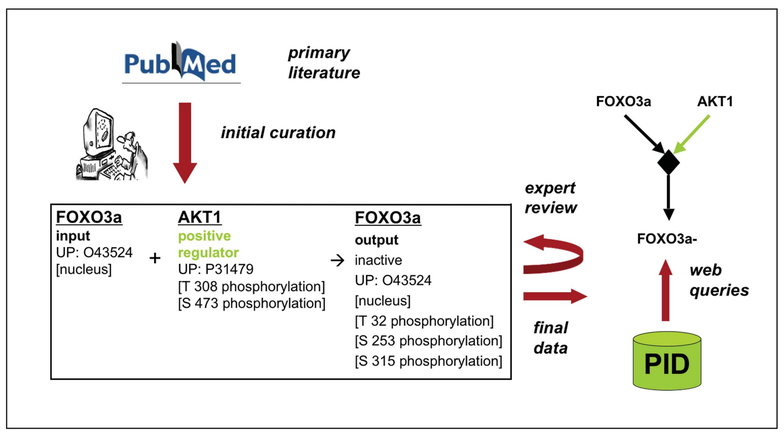
Intra-pathway assembly
Role of biomolecule (edge): Each biomolecule can be an **input**, **output**, **positive regulator** or **negative regulator**. Inputs are transformed into outputs, and regulators act either directly or indirectly on the input. An output biomolecule can form the input, positive regulator or negative regulator biomolecule in subsequent interactions.
Generic process types (nodes): Four process types describe the general nature of each interaction. These process types are: **Modification** (binding) including post-translational modifications, **translocation**, **transcription** and **reaction**. Each process may also possess a condition, which is a requirement for the interaction to take place. A condition may be defined by a GO Biological Process term, a GO Molecular Function term or the NCI Thesaurus if a cell type-specific condition is necessary.
Descriptive process types (node): **Macroprocesses** are multi-step events that are often defined by a GO Biological process term (e.g. activated T cell apoptosis) or by the NCI Thesaurus if a cancer-related term is necessary.

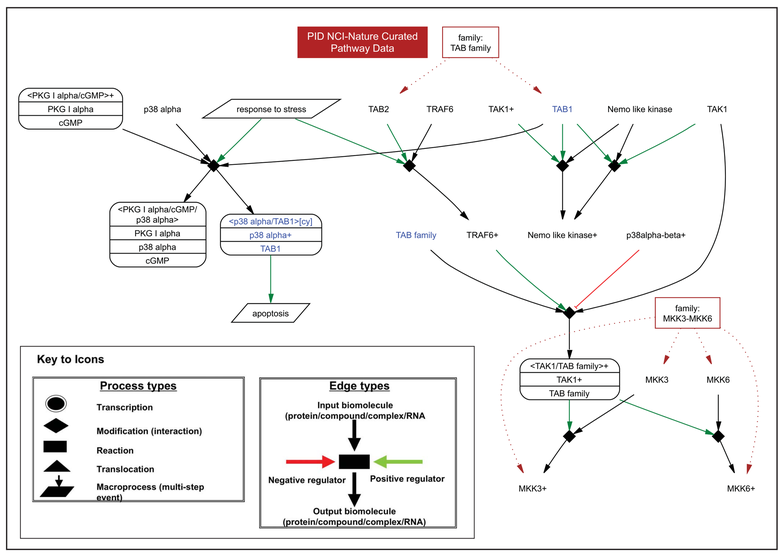
Evidence
Interactions are supported by references, annotated with PubMed identifiers. Evidence codes are also included in the interaction, and are partially derived from GO Evidence codes. PID evidence codes are as follows:

Acronym	Expansion	Description
IAE	Inferred from Array Experiments	ChIP on chip Protein microarrays/protein chips Chemical compound arrays
IC	Inferred by Curator	An experimentally-determined interaction, not falling under any of the other evidence codes, but still deemed to occur by the curator. (Not commonly applied.)
IDA	Inferred from Direct Assay	Enzyme assays In vitro reconstitution Functional and activity assays
IFC	Inferred from Functional Complementation	A gene from one organism complements a mutation in another species.
IGI	Inferred from Genetic Interaction	Genetic interactions such as suppressors, synthetic lethals, and rescue experiments Inference about one gene drawn from the phenotype of a mutation in a different gene
IMP	Inferred from Mutant Phenotype	Any gene mutation/knockout Overexpression/ectopic expression of wild-type or mutant genes Anti-sense experiments RNAi experiments Polymorphism or allelic variation
IOS	Inferred from Other Species	An interaction that is inferred from another species due to a lack of evidence in human.
IPI	Inferred from Physical Interaction	Any physical interaction detection method, common ones are: 2-hybrid interactions co-purification, co-immunoprecipitation, and ion/protein binding experiments.
RCA	Inferred from Reviewed Computational Analysis	Predictions based on large-scale experiments (e.g. genome-wide two-hybrid, genome-wide synthetic interactions) Predictions based on integration of large-scale datasets of several types Text-based computation (e.g. text mining)
RGE	Inferred from Reporter Gene Expression	Reporter gene expression studies
TAS	Traceable Author Statement	Anything found in secondary literature (such as a review article or textbook) where the original experiments are traceable through that piece of literature.

Inter-pathway assembly
To facilitate additional connectivity, curated pathways may be linked to from one another. In addition, to arrange the cascade of events, pathways may be divided into smaller, biologically meaningful **subnetworks**.

PID homepage


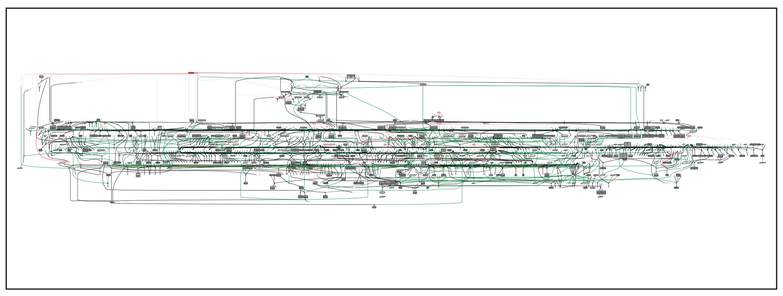
Curation process


Interactive network map


Predefined pathways and dynamically generated network maps based on web queries are generated by the freely available GraphViz (**http://www.graphviz.org**) program and can be visualized in either Scalable Vector Graphics (SVG) or Graphics Interchange (GIF) Formats. The SVG format requires that a plugin be downloaded from **http://www.adobe.com/SVG/viewer/install/main.html**. The network maps are interactive; users can click on a biomolecule or interaction icon for more information. The search term, TAB1, is shown in blue.

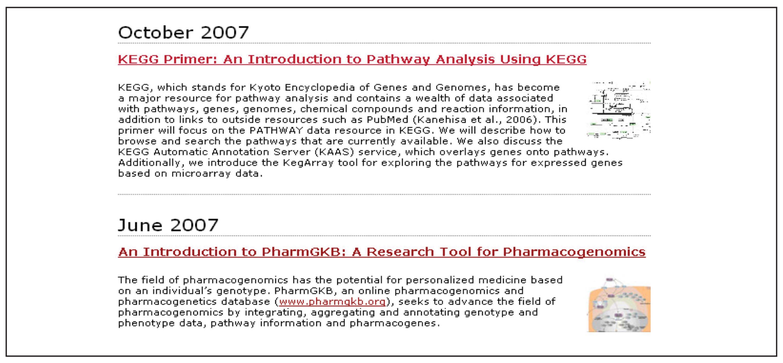
Browsing & searching
Users can browse complete, predefined pathways, and perform simple or advanced molecule or process searches. Searches can also be limited by evidence code.

Batch query
The Batch query allows users to upload long lists of molecules and analyze their relationships in pathways. Two lists can be uploaded simultaneously in order to compare data sets (e.g. gene expression results). Using the list from Stratton and colleagues (2007) of 120 protein kinases found to contain at least one cancer-predisposing mutation and selecting the NCI-Nature curated data source, users can view the network connectivity of the uploaded molecules. The output is a set of 19 distinct network maps (the one with the most interactions is shown below). Query molecules are shown in blue in the network map:



Editorial content
Research highlights
Relevant cell signaling and bioinformatics research highlights written by the Nature Reviews team are featured in the PID monthly updates.

Bioinformatics primers
Bioinformatics primers provide practical advice on how to make the most of important online resources. Previously published Bioinformatics primers include:



We are currently commissioning Bioinformatics primers for the second half of 2008.

Data download options
Users can download data in either XML (Extensible Markup Language) or BioPAX (**http://www.biopax.org/**) formats by visiting the Download data page at **http://pid.nci.nih.gov/PID/download.shtml**. The PID supports BioPAX Level 2 format for pathway data exchange, which includes metabolic pathways, molecular interactions and protein post-translational modifications.

Community
BioPAX
The PID is participating in the development of future levels of BioPAX, which will increase support for signaling pathways, gene regulatory networks and genetic interactions.

Pathway Commons
PID pathways are also featured in Pathway Commons (**http://www.pathwaycommons.org**), which is a biological pathways portal allowing for pathway visualization and analysis via Cytoscape (**http://www.cytoscape.org**).

References
An Introduction to the NCI-Nature Pathway Interaction Database
Carl F. Schaefer

Pathway Interaction Database (09 November 2006) DOI: 10.1038/PID.2006.001

Patterns of somatic mutation in human cancer genomes.
Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR

Nature. 2007;446;153-8. PMID: 17344846 DOI: 10.1038/nature05610

Nature Precedings : doi:10.1038/npre.2007.1351.1 : Posted 12 Nov 2007