Published online 19 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D793–D799***
doi:10.1093/nar/gkq1167

SPIKE: a database of highly curated human signaling pathways

Arnon Paz[1], Zippora Brownstein[1], Yaara Ber[2], Shani Bialik[2], Eyal David[1], Dorit Sagir[1], Igor Ulitsky[3], Ran Elkon[1], Adi Kimchi[2], Karen B. Avraham[1], Yosef Shiloh[1] and Ron Shamir[3,*]

[1]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, [2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot and [3]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Received August 10, 2010; Revised October 29, 2010; Accepted November 1, 2010

ABSTRACT

The rapid accumulation of knowledge on biological signaling pathways and their regulatory mechanisms has highlighted the need for specific repositories that can store, organize and allow retrieval of pathway information in a way that will be useful for the research community. SPIKE (Signaling Pathways Integrated Knowledge Engine; http://www.cs.tau.ac.il/~spike/) is a database for achieving this goal, containing highly curated interactions for particular human pathways, along with literature-referenced information on the nature of each interaction. To make database population and pathway comprehension straightforward, a simple yet informative data model is used, and pathways are laid out as maps that reflect the curator's understanding and make the utilization of the pathways easy. The database currently focuses primarily on pathways describing DNA damage response, cell cycle, programmed cell death and hearing related pathways. Pathways are regularly updated, and additional pathways are gradually added. The complete database and the individual maps are freely exportable in several formats. The database is accompanied by a stand-alone software tool for analysis and dynamic visualization of pathways.

INTRODUCTION

The dynamic behavior of biological systems and their response to various stimuli and physiological changes are driven by signaling networks mobilized primarily by alterations in gene and protein activity. Identifying the signaling pathways and specific regulation mechanisms is a major effort in biomedical research. Many of these regulations have been identified individually using *ad hoc* techniques. More recently, high-throughput techniques are rapidly adding new information by generating data on a larger scale albeit at lower fidelity (1–5). The volume of this information is such that even for a single pathway, it is difficult to recall all the interactions involved and the experimental source of each piece of information. SPIKE (Signaling Pathways Integrated Knowledge Engine; http://www.cs.tau.ac.il/~spike/) aims to construct, archive and actively maintain a database of highly curated interactions for particular human pathways, along with information on the nature of each interaction and its reference in the literature. High curation quality is obtained by supervision of map creation by leading domain experts. The pathways are laid out as maps that reflect the curator's understanding and intuition and make pathway utilization straightforward. The current focus of the database is primarily on pathways describing DNA damage response, cell cycle, programmed cell death and hearing related pathways, and additional pathways are being added continuously. The database is accompanied by a stand-alone software tool for analysis and dynamic visualization of pathways (6). Here we focus on the structure and contents of the database.

There are several important databases that collect and curate protein interactions and pathways. Among those are KEGG Pathways (7), Reactome (8), ConsensusPathDB (9), IntAct (10) and NetPath (11). SPIKE differs from some of these in the choice of data model, emphasizing a simple model using a single entity per gene/protein in the underlying database, in its focus on signaling regulations (and excluding others, e.g. metabolic reactions), and in the continuous update of focus maps.

*To whom correspondence should be addressed. Tel: +972 3 640 5383; Fax: +972 3 640 5384; Email: rshamir@tau.ac.il
Present addresses:
Igor Ulitsky, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA.
Ran Elkon, Division of Gene Regulation, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

It is unique in the focus on specific pathways of expertise, for which highly curated information is provided and regularly updated by domain experts. In fact, due to the differences in data model and curation policy, maps of the same process in SPIKE and other databases may manifest marked dissimilarities, with consequences for downstream analysis (see Supplementary Table S1 for comparison between few SPIKE and KEGG maps which cover the same pathway/process). The simple model and intuitive maps allow new maps to be readily constructed by experts in a distributed community effort. Like many other databases, it imports data from other databases and also allows freely export of the data in commonly used formats, such as BioPax (12).

One of the main utilities of pathway databases is in analysis pipelines of 'omics' data sets. Such data sets provide genome/transcriptome/proteome-wide snapshots of cellular processes and mining meaningful biological insights out of them poses a major challenge. To cope with this task we have integrated SPIKE maps in the EXPANDER package which is developed in our lab (13). The availability of SPIKE maps in standard formats readily allows their integration in other 'omics'-analysis tools.

While SPIKE imports data and pathways en masse from other databases, it emphasizes completeness of its core pathways, performing continuous updates based on recent reports in the literature. All interactions are stored in the database with reference information, and pathways are graphically laid out by experts for best comprehension by users. Those interested in modifying layouts and exploring the pathway 'boundaries' using additional data stored in the database can download the software tool and then manipulate and explore the subnetworks in many ways.

DATA MODEL

The need to represent biological knowledge in a formal language within electronic knowledge-bases is well recognized and several ontologies have been defined [e.g. (14)]. We envisioned and implemented the SPIKE database as a community tool whose 'upper tier' contains highly-curated data, contributed by experts in various domains in the biomedical research (see Supplementary Data for a description of our curation policy). To allow swift and easy database population in a distributed fashion and to keep SPIKE maps easy to comprehend, we adopted model simplicity as a major design principle. Accordingly, the data model focuses on fundamental features that characterize regulatory events that build signaling networks while deliberately ignoring, for the sake of simplicity, other aspects of such events, which we deem less crucial. The data model includes five types of biological entities (nodes in SPIKE maps) and three types of relationships between the entities (edges in the maps) (Figure 1A).

The five types of biological entities in the data model are: (i) genes (and their RNA and, in case of protein-coding genes, protein products) form the elementary object in the model. To avoid ambiguities, only human genes assigned with a formal designation by the Human Genome Nomenclature Committee (HGNC) are included in SPIKE's gene space and are uniquely identified by their Entrez Gene IDs (15). For simplicity, we decided not to define distinct objects for genes, their transcribed RNAs and translated proteins. (ii) Families are groups of isoform genes (encoded by distinct genomic loci) with high sequence homology and whose encoded transcripts/proteins share most of their biological activities. (iii) Complexes are groups of proteins that carry out a specific function only when associated with their complex mates. The complex entity supports a nested structure, namely, a complex may contain sub-complexes. (iv) Chemical molecules which participate in regulatory networks (e.g. cAMP, Ca^{++}, GTP) are identified in SPIKE database by their ID in ChEBI (16). (v) General entity. This type was added to increase the model flexibility and allow users to add to maps nodes not covered by the first four types. This type of nodes is most often but not exclusively used to describe biological processes or events that are either triggers or endpoints of the signaling pathways described in the respective maps.

The types of relationships between SPIKE entities are: (i) containment defines relationships between families/complexes and the members they contain. (ii) Regulation defines a directed and signed regulatory link between its source and target entities. Each regulation is defined by a source, a target and effect. It is also associated with several additional attributes: the biochemical mechanism by which it is driven (e.g. phosphorylation, transcriptional regulation), one or more supporting references, quality level for quality control (see the Supplementary Data), and the submitter/data source. (iii) Interaction defines an undirected and unsigned physical link between two protein nodes. Each interaction is defined by its two participating proteins, and also has an attribute that indicates the experimental method used to identify it (e.g. Y2H, co-immunoprecipitation) in addition to attributes indicating supporting reference(s), quality and the data source.

DATA SOURCES AND EXCHANGE WITH OTHER DATABASES

SPIKE database imports data from multiple sources (Figure 1B). Genes in SPIKE database are imported from Entrez Gene (15) (we include in SPIKE only genes with 'Reviewed' RefSeq status) and chemical molecules are imported from ChEBI (16). SPIKE's data on relationships between entities come from three sources: (i) Highly curated data submitted directly to SPIKE database by SPIKE curators and experts in various biomedical domains. (ii) Data imported from external signaling pathway databaes. At present, SPIKE database imports such data from Reactome (8), KEGG (7), NetPath (11) and The Transcription Factor Encyclopedia (http://www.cisreg.ca/cgi-bin/tfe/home.pl). (iii) Data on protein–protein interactions (PPIs) imported either directly from wide-scale studies that recorded such interactions [to date,

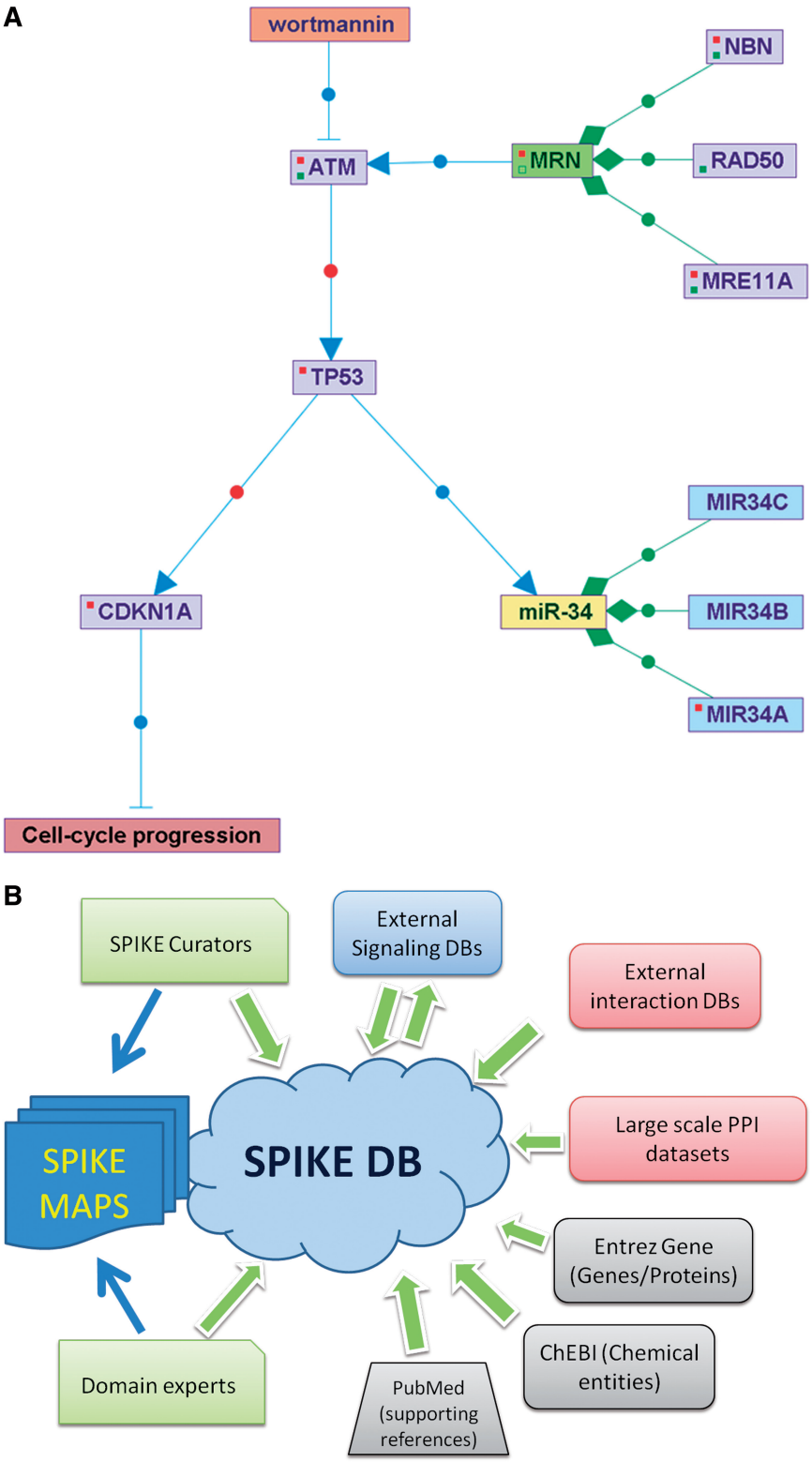

Figure 1.SPIKE's data model and links with other databases. (A) The data model. SPIKE's data model includes five types of biological entities (nodes in SPIKE maps) and three types of relationships between entities (edges in the maps). The types of entities are: (i) Genes/Proteins. Protein-coding genes are displayed in the maps as violet (e.g. ATM in the figure); non-coding genes are light-blue nodes (e.g. MIR34C). (ii) Families (yellow nodes, e.g. MIR-34), (iii) Complexes (green nodes, MRN), (iv) Chemical molecules (orange nodes, wortmannin) and (v) General entities (dark-pink nodes; e.g. 'cell-cycle progression' in this map). The types of relationships are: (i) Containment links between families or complexes and their members, shown as green edges (e.g. miR34B is contained in the MIR-34 family). (ii) Regulations, displayed as directed blue edges; arrows represent positive regulation (e.g. ATM activates TP53) and T-shape edges indicate negative regulation (e.g. wortmannin inhibits ATM). (iii) Interactions shown as undirected blue edges (not included in this figure). Red/green dots within a node indicate that the node has additional regulations/containments in the SPIKE database that are not included in the map. The dots on the edges can be used in the SPIKE stand-alone version to identify the literature reference to the relationship and as a

Continued

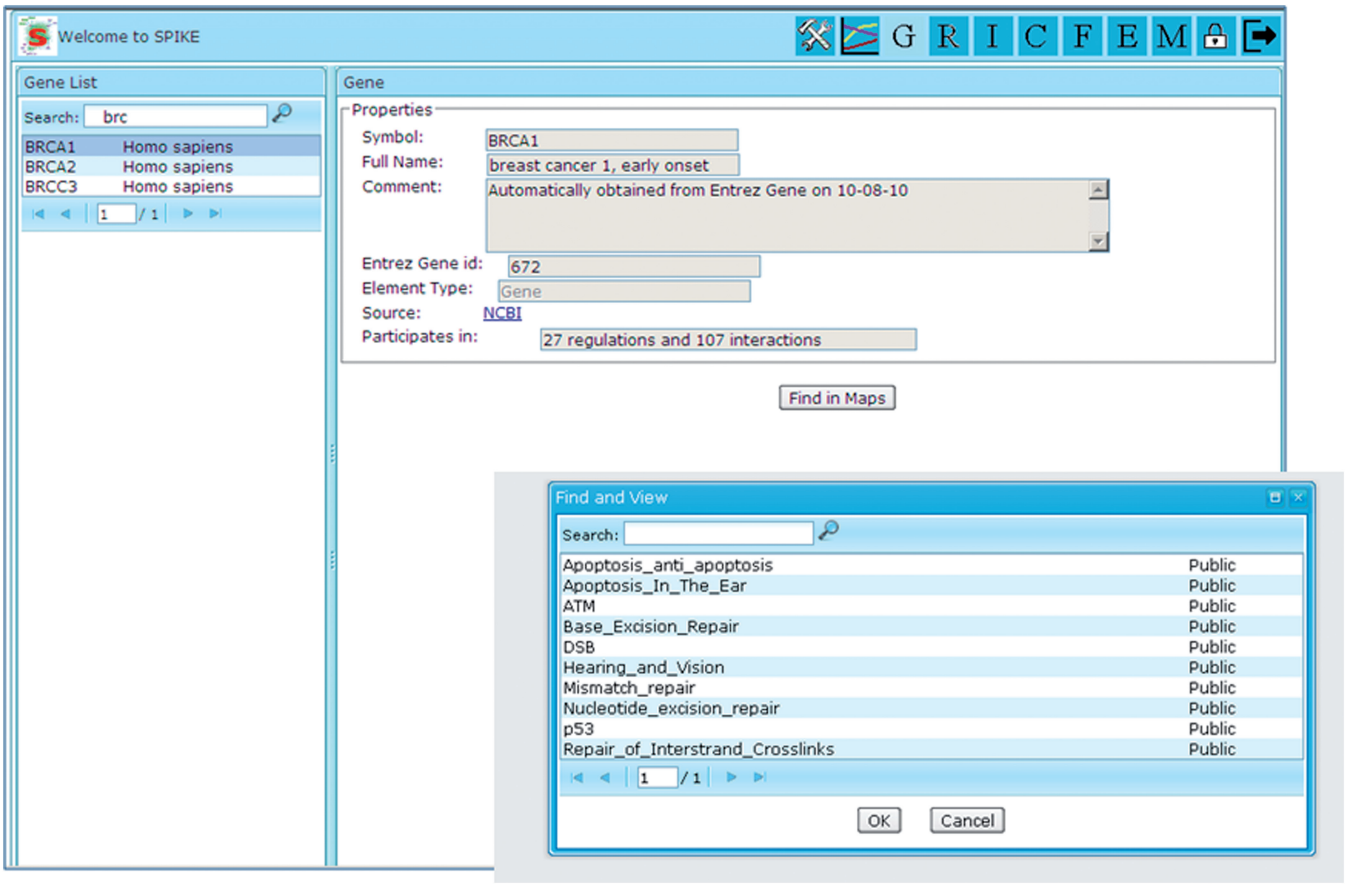

Figure 2. Browsing SPIKE database. Basic searching utilities over SPIKE database are provided in the website. In this example the BRCA1 gene was searched. Pertinent information about the gene and a list of the maps in which it is included are displayed to the user.

PPI data were imported from Stelzl *et al.* (17), Rual *et al.* (3) and Lim *et al.* (18)] or from external PPI databases [IntAct (10) and MINT (19)]. Relationship data coming from these different sources vary greatly in their quality and this is reflected by a quality level attribute, which is attached to each relationship in SPIKE database (Supplementary Data). Each relationship in SPIKE is linked to at least one PubMed reference that supports it.

In addition to the individual maps, the complete database is freely available for download in several exchange formats. See under 'Data availability' below.

BROWSING THE DATABASE

SPIKE website offers basic searching utilities over the SPIKE database. Users can search the databases for each of the basic building blocks of SPIKE data model (namely, the five types of biological entities and the three types of relationships). For each entity, pertinent information is displayed (e.g. full name, description and external-ID for genes; data source, supporting PubMed references and quality level for regulations) and a list of the maps that contain the inquired entity is shown (Figure 2). The same interface serves SPIKE curators and registered users for uploading data to the database.

DATABASE CONTENTS

SPIKE-database was greatly expanded since our previous publication (6). The number of highly curated regulations has increased 8-fold, and the number of maps has increased from less than 10 to 23. In addition, database web-browsing capabilities were added. As of August 2010, the SPIKE database contains 20 412 genes/proteins, 542

Continued
handle to move the edge center in the map for better visualization. (**B**) SPIKE data sources and interlinks with external databases. The upper tier of SPIKE database contains highly curated data either uploaded by SPIKE curators or contributed by experts in various domains in the biomedical research. Each regulation uploaded by the curators or experts is supported and linked to at least one reference (PubMed). In addition, SPIKE imports data from external signaling databases and protein–protein databases and large-scale data sets. Data uploaded to SPIKE database receive quality flag that indicate their reliability level (see text for details); in this figure data sources are color-coded according to the quality flag they are assigned to: Data uploaded by SPIKE curators and domain experts get the highest quality level; Data imported from external curated signaling databases are assigned intermediate quality level (since these data are imported automatically and en masse into SPIKE database) and data imported from protein–protein interaction databases and data sets get low quality level. SPIKE genes and chemical molecules are imported from Entrez-Genes and ChEBI, respectively.

Table 1. SPIKE maps

Map	Nodes (genes, complexes, families)	Links (regulations, interactions)	Unique references used	Creation date	Last update
Cell cycle progress and check points					
G1-S Phase	55	70	28	August 2006	January 2007
G2-M Phase	78	122	90	August 2006	January 2010
DNA damage response					
Response to double strand breaks	144	232	144	January 2009	December 2009
Nucleotide excision repair	94	136	88	October 2009	
ATM signaling network	118	152	120	January 2009	August 2010
Repair of Interstrand Crosslinks	86	139	99	January 2010	
Base Excision Repair (BER)	82	120	79	March 2010	August 2010
Mismatch repair (MMR)	50	72	48	March 2010	
Programmed cell death related processes					
Apoptosis	94	217	168	January 2009	April 2010
Caspases Cascade	113	216	151	October 2009	
DAPk family	69	93	77	October 2009	
Apoptosis Anti-Apoptosis Network	120	232	165	January 2009	August 2009
Autophagy	57	93	93	July 2009	
Stress-activated transcription factors					
p53 Signaling Network	91	111	88	August 2006	August 2009
NFkB Signaling Network	89	123	137	August 2006	August 2009
Mitogen-activated protein kinase pathways					
MAPK signaling	76	89	49	September 2006	August 2009
Immune response signaling					
TLR Signaling	134	187	110	January 2007	March 2010
HEarSpike: hearing related pathways					
Hearing related SIX1 Interaction	103	131	79	June 2009	
MYO7A Interactions In The Ear	62	76	38	June 2009	
NOTCH1 Signaling In The Ear	76	93	53	June 2009	
Apoptosis in the ear	178	279	195	March 2010	
Hearing and vision	258	457	320	March 2010	
NYO3A	98	128	99	August 2010	

complexes (327 of high quality), 320 protein families (167 of high quality) and 39 small molecules. These entities are linked by 34 338 interactions (of which 2400 are of high quality) and 6074 regulations (4420 of high quality). These are associated with 5873 journal references in total. There are 23 different maps covering the areas of DNA damage responses, programmed cell death, hearing-related pathways, cell cycle regulation and more (Table 1). The database is undergoing rapid growth, and within the last 2 years the number of high quality regulations and the number of maps more than doubled. A brief description of the main domains of focus, and the maps created for each one is provided in the Supplementary Data. An example of the autophagy map is shown in Figure 3.

USER INTERFACE

In addition to the directly downloadable database, SPIKE provides a stand-alone software package for visualization and analysis of the database and maps. The software is available for download from SPIKE site (http://www.cs.tau.ac.il/~spike/download.html). The installation package contains the complete database with fully updated data. The visualization package allows interactive graphic representations of regulatory interactions stored in the database and superposition of functional genomic and proteomic data on the maps. The software also includes an algorithmic inference engine that analyzes the networks for novel functional interplays between network components. Interconnection with analysis and visualization tools is under development (Supplementary Data).

Accessing public maps with SPIKE software enables viewing the maps with the same layout and visual properties as they were constructed and allows the user to edit them. Furthermore once a map is opened using the tool, every object in the map is scanned for updates in the SPIKE database. All new regulations and interactions that were updated in the database since the map was created are conveniently arranged to be added or ignored. The software can save or convert the edited maps to SIF or BioPax formats.

DATA AVAILABILITY

An up-to-date snapshot of the SPIKE database is freely available for download from the SPIKE site (http://www.cs.tau.ac.il/~spike/download/LatestSpikeDB.xml.zip). Currently downloads can be made in three formats: BioPax, SPIKE's unique XML format, and SIF. SPIKE's specific XML format provides a copy all the information contained in the database (nodes, regulations and interactions, with associated links and literature reference information on each entity) in a format readable by the SPIKE software. The Supplementary Data describes the format and shows an example of the result of exporting a small map to XML and presenting it with Cytoscape (20). The SIF snapshot contained limited information on

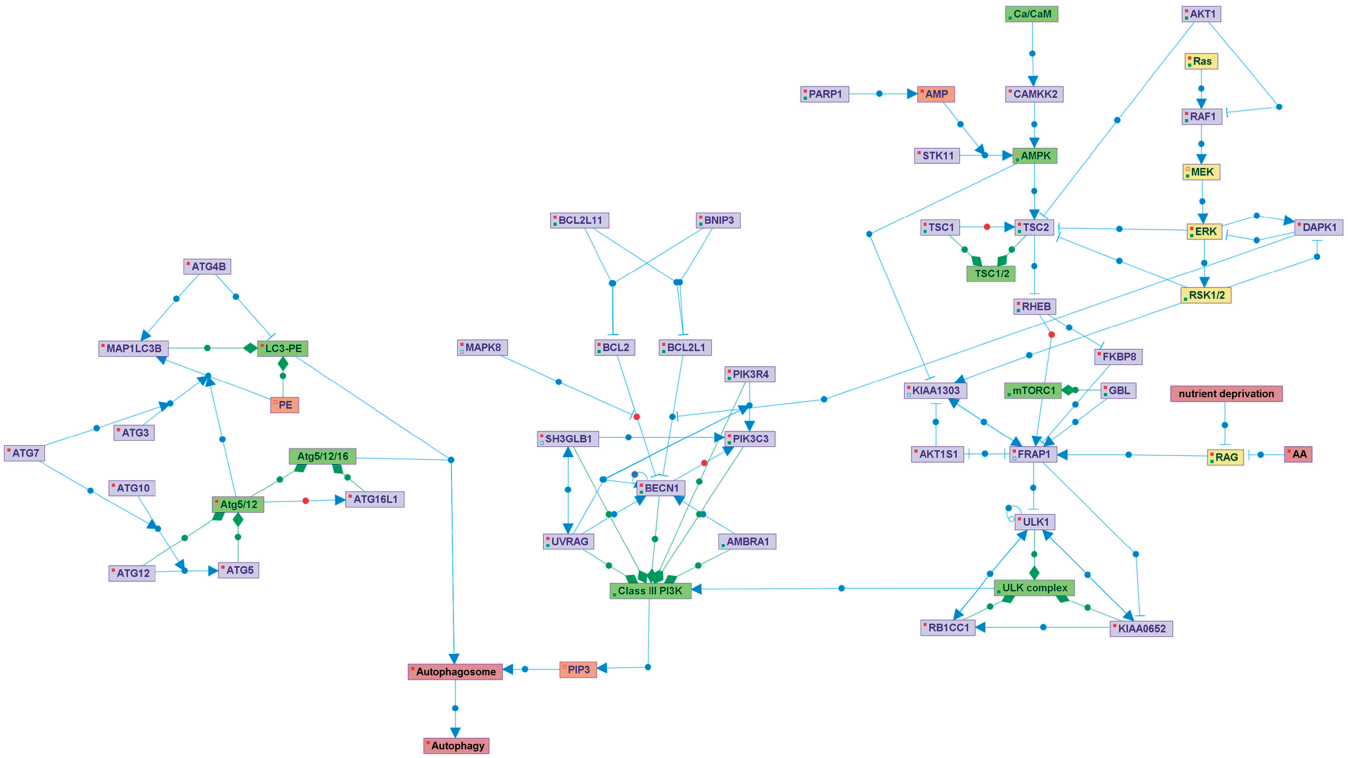

Figure 3. Autophagy map. Autophagy, or 'self eating', is the process by which de novo-formed double-membrane enclosed vesicles, known as autophagosomes, engulf cytosolic components and organelles, and ultimately fuse with the lysosome, enabling degradation of its internal contents. Autophagy serves as a means of recycling critical cellular building blocks, especially in times of deprivation and stress, and removal of damaged organelles, and misfolded and aggregated proteins. The main regulators and effectors of this process and interlinks between them are depicted in the map. The induction of autophagy is modulated by several signaling pathways which converge at the activation of the mTOR/ULK complex (right part of the map) and of Class III PI3K complex (center). Two ubiquitin-like pathways (left part of the map) mediate the elongation of autophagosome membranes.

interactions since this format is not rich enough include all the information. It is useful primarily for constructing graph views of the map using other software tools.

Individual SPIKE curated maps are downloadable from the website as well. The maps are provided in the same formats as the database snapshot. Maps in SPIKE's XML format are editable using the SPIKE software tool. Maps in SIF and BioPax format can be viewed using network visualization tools such as Cytoscape. The visualization of the maps in these formats will look different than it appears on the SPIKE site since graph layout information is not supported by SIF and BioPax.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

TRIREME project, EC Seventh Framework Programme (grants HEALTH-F4-2009-223575 to R.S. and Y.S.); HEALTH-F4-2007-200767 for the APO-SYS project (to R.S. and A.K.); Converging Technologies Program of the Israel Science Foundation (grant 1767.07 to R.S. and Y.S.); A-T Medical Research Foundation (to Y.S.); the Israel Cancer Research Fund (to Y.S.); The Dr Miriam and Sheldon G. Adelson Medical Research Foundation (to Y.S.); Israel Science Foundation (grant 1486/07 to K.B.A.); NIH/NIDCD (grant R01 DC005641 to K.B.A.); European Commission FP6 Eumodic 037188 (to K.B.A.); Center of Excellence grant from the Flight Attendant Medical Research Institute (FAMRI, to A.K.). A.K. is the incumbent of Helena Rubinstein Chair of Cancer Research. Y.S. is a Research Professor of the Israel Cancer Research Fund. R.S. is the incumbent of the Raymond and Beverly Sackler Chair in Bioinformatics; Edmond J. Safra Bioinformatics program at Tel-Aviv University, fellowship (A.P. and I.U.). Funding for open access charge: TRIREME project, EC Seventh Framework Programme (grants HEALTH-F4-2009-223575 to R.S. and Y.S.).

Conflict of interest statement. None declared.

REFERENCES

1. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
2. Braun,P., Tasan,M., Dreze,M., Barrios-Rodiles,M., Lemmens,I., Yu,H., Sahalie,J.M., Murray,R.R., Roncari,L., de Smet,A.S. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods*, **6**, 91–97.

3. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
4. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
5. Berns,K., Hijmans,E.M., Mullenders,J., Brummelkamp,T.R., Velds,A., Heimerikx,M., Kerkhoven,R.M., Madiredjo,M., Nijkamp,W., Weigelt,B. *et al.* (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*, **428**, 431–437.
6. Elkon,R., Vesterman,R., Amit,N., Ulitsky,I., Zohar,I., Weisz,M., Mass,G., Orlev,N., Sternberg,G., Blekhman,R. *et al.* (2008) SPIKE: a database, visualization and analysis tool of cellular signaling pathways. *BMC Bioinformatics*, **9**, 110.
7. 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.
8. 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.
9. Kamburov,A., Wierling,C., Lehrach,H. and Herwig,R. (2009) ConsensusPathDB: a database for integrating human functional interaction networks. *Nucleic Acids Res.*, **37**, D623–D628.
10. Aranda,B., Achuthan,P., Alam-Faruque,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.
11. Kandasamy,K., Mohan,S.S., Raju,R., Keerthikumar,S., Kumar,G.S., Venugopal,A.K., Telikicherla,D., Navarro,J.D., Mathivanan,S., Pecquet,C. *et al.* (2010) NetPath: a public resource of curated signal transduction pathways. *Genome Biol.*, **11**, R3.
12. 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.
13. Ulitsky,I., Maron-Katz,A., Shavit,S., Sagir,D., Linhart,C., Elkon,R., Tanay,A., Sharan,R., Shiloh,Y. and Shamir,R. (2010) Expander: from expression microarrays to networks and functions. *Nat. Protoc.*, **5**, 303–322.
14. Fukuda,K. and Takagi,T. (2001) Knowledge representation of signal transduction pathways. *Bioinformatics*, **17**, 829–837.
15. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
16. de Matos,P., Alcántara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
17. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
18. Lim,J., Hao,T., Shaw,C., Patel,A.J., Szabo,G., Rual,J.F., Fisk,C.J., Li,N., Smolyar,A., Hill,D.E. *et al.* (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, **125**, 801–814.
19. Ceol,A., Chatr Aryamontri,A., Licata,L., Peluso,D., Briganti,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
20. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.