D470–D478 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 11 October 2019*
doi: 10.1093/nar/gkz861

PathBank: a comprehensive pathway database for model organisms

David S. Wishart[1,2,*], Carin Li[1], Ana Marcu[1], Hasan Badran[1], Allison Pon[1], Zachary Budinski[1], Jonas Patron[1], Debra Lipton[1], Xuan Cao[1], Eponine Oler[1], Krissa Li[1], Maïlys Paccoud[1], Chelsea Hong[1], An C. Guo[1], Christopher Chan[1], William Wei[1] and Miguel Ramirez-Gaona[3]

[1]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada, [2]Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada and [3]Department of Plant Breeding, Wageningen University and Research, 6708 PB Wageningen, Gelderland, Netherlands

Received August 15, 2019; Revised September 14, 2019; Editorial Decision September 21, 2019; Accepted October 08, 2019

ABSTRACT

PathBank (www.pathbank.org) is a new, comprehensive, visually rich pathway database containing more than 110 000 machine-readable pathways found in 10 model organisms (*Homo sapiens, Bos taurus, Rattus norvegicus, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae, Escherichia coli* and *Pseudomonas aeruginosa*). PathBank aims to provide a pathway for every protein and a map for every metabolite. This resource is designed specifically to support pathway elucidation and pathway discovery in transcriptomics, proteomics, metabolomics and systems biology. It provides detailed, fully searchable, hyperlinked diagrams of metabolic, metabolite signaling, protein signaling, disease, drug and physiological pathways. All PathBank pathways include information on the relevant organs, organelles, subcellular compartments, cofactors, molecular locations, chemical structures and protein quaternary structures. Each small molecule is hyperlinked to the rich data contained in public chemical databases such as HMDB or DrugBank and each protein or enzyme complex is hyperlinked to UniProt. All PathBank pathways are accompanied with references and detailed descriptions which provide an overview of the pathway, condition or processes depicted in each diagram. Every PathBank pathway is downloadable in several machine-readable and image formats including BioPAX, SBML, PWML, SBGN, RXN, PNG and SVG. PathBank also supports community annotations and submissions through the web-based PathWhiz pathway illustrator. The vast majority of PathBank's pathways (>95%) are not found in any other public pathway database.

INTRODUCTION

Pathway diagrams are the roadmaps for molecular biology. Just as roadmaps show the connections between villages, towns, and cities, pathway diagrams illustrate the connections between genes, proteins and metabolites. A well-illustrated pathway diagram provides biological context to complex molecular processes in an easily understood and highly visual manner. In this regard, pathway diagrams provide a remarkably useful means for scientists to share, integrate, interpret and visualize ‘omics data and ‘omics measurements. Even now, compilations of pathway diagrams (i.e. pathway databases) are among the most frequently used and oft-referenced resources in the entire field of bioinformatics. This is because pathway databases are widely called upon to interpret genomics, metagenomics, transcriptomics, proteomics and/or metabolomics data. Furthermore, they are frequently used to facilitate more complex analytical tasks such as cellular modeling (1), flux balance analysis (2), gene-set enrichment analysis (3), metabolite-set enrichment analysis (4) and over-representation analysis (5).

A wide variety of pathway databases exist, including both commercial and open-access resources. Some of the best-known online pathway databases are also among the oldest. These include KEGG (6) and the ‘Cyc’ databases (7), both of which emerged in the mid 1990s. Both of these resources are superb collections, providing extensive, easily searched, web-accessible pathways and annotations that are widely used by thousands of scientists each day. More recently, a number of other pathway databases have emerged such as Reactome (8) and Wikipathways (9), which cover pathways for a selected (small) number of model organisms.

[*]To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: david.wishart@ualberta.ca

Additionally, Wikipathways employs a community annotation model and exploits a number of continuing advances in web visualization and interactivity to offer a somewhat richer user/contributor experience. There are also a variety of specialized pathway databases such as BioCarta (https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) (10), InnateDB (11), PharmGKB (12) and the Pathway Interaction Database (13), which either focus on specific pathway types, specific (single) organisms or provide a different, often visually more impressive way of exploring pathways than their comprehensive pathway database cousins.

In addition to these freely available or open-access pathway databases, there are also several well-known commercial pathway databases such as GeneGo (https://portal.genego.com/), Protein Lounge (http://www.proteinlounge.com/) or IPA (Ingenuity Pathway Analysis) (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/). In particular, the Ingenuity Pathway Analysis system by Qiagen is perhaps the most popular and most extensive commercial resource in terms of pathway coverage and pathway analysis tools.

While there are many excellent features in each of the pathway databases just mentioned, there are also some notable limitations as well. For instance, even though KEGG and the Cyc databases are explicitly small molecule pathway databases, they do not provide many pathways covering lipid synthesis, metabolite signaling, small molecule hormone signaling or small molecule drug action. While the KEGG database does provide some protein signaling, developmental and disease process pathways, the CycDBs do not. Neither KEGG nor the CycDBs provide pathways relating to cellular responses, and KEGG does not provide information on cellular locations or biological context. Being the oldest of today's pathway databases, KEGG and the Cyc databases also use somewhat dated approaches to web-based pathway visualization. While the Reactome and Wikipathways databases do provide somewhat more modern web interfaces and cover many aspects of protein and cellular signaling, they are also much more limited in their coverage of metabolism than KEGG or the Cyc databases. Likewise, Reactome and Wikipathways largely ignore metabolite signaling, hormone signaling, drug action and many common disease pathways. Even with their modern pathway visualization tools, the Reactome and Wikipathway pathway diagrams still lack the visual appeal or biological context of the pathway diagrams found in other pathway databases such as BioCarta (10), DrugBank (14) or SMPDB (15).

Ideally what is needed is a pathway database that combines the comprehensive metabolic pathway coverage found in the KEGG or the Cyc databases, with the rich protein/cellular signaling pathway coverage found in Reactome, Wikipathways or IPA, and the visualization quality and biological context found in BioCarta or SMPDB—into a single, freely available resource. Herein we describe just such a database—called PathBank. PathBank is a comprehensive, visually rich pathway database containing more than 110 000 machine-readable pathways found in 10 model organisms (including humans, rodents, plants, insects and microbes). PathBank provides detailed, fully searchable, hyperlinked diagrams of metabolic, protein/metabolite signaling, disease, drug and physiological pathways. Its pathway diagrams include information on the relevant organs, organelles, subcellular compartments, locations and structures to provide important biological context. All molecular/protein images are hyperlinked to databases with rich descriptions and all pathways are viewable using a modern 'Google Maps'-like or AJAX-based visual navigation system. All of PathBank's pathways have detailed descriptions and all of its images, image maps, descriptions and reactions are freely downloadable in several common data exchange formats. The entire database may be browsed, searched or annotated using a variety of easy-to-use tools. Likewise, community or crowd-sourced annotations and submissions to PathBank are supported through the web-based PathWhiz (16) pathway illustrator.

PathBank DESCRIPTION AND CONTENT

PathBank (version 1.0) contains 110 234 richly annotated, fully colored, machine-readable pathways for 10 model organisms (*Homo sapiens, Bos taurus, Rattus norvegicus, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae*, *Escherichia coli* and *Pseudomonas aeruginosa*). The database currently contains 78 488 different compounds (including metabolites, drugs and other xenobiotics), 8993 different proteins and describes more than 176 535 different reactions and interactions. PathBank pathways are divided into two broad categories or classes: (i) Metabolite/Compound pathways and (ii) Protein pathways. Metabolite/Compound pathways contain a majority (>50%) of small molecules compared to proteins as pathway entities. Protein pathways contain a vast majority of proteins (typically >80%) compared to metabolites as pathway entities. In the current version of PathBank there are 109 836 Metabolite/Compound pathways and 398 Protein pathways.

Within PathBank there are six subcategories of Metabolite/Compound pathways: (i) Metabolic (catabolism/anabolism); (ii) Physiological/Endocrinological (primarily involving metabolites); (iii) Metabolite Signaling; (iv) Drug Metabolism; (v) Drug Action; and (vi) Disease (primarily involving small molecule metabolism). In contrast to the Metabolite/Compound pathways there are 15 subcategories of Protein pathways: (i) Immunological; (ii) Cellular Response; (iii) Gene Regulatory; (iv) Growth Factor; (v) Cytokine Signaling; (vi) Protein/Peptide Hormone-Mediated; (vii) Neurological Signaling; (viii) Developmental Signaling; (ix) Kinase Signaling; (x) Apoptosis Signaling; (xi) Stress Activated Signaling; (xii) Pathogen-Activated Signaling; (xiii) Transport/Degradation; (xiv) Cytoskeletal Signaling; and (xv) Disease (primarily involving protein dysregulation).

All pathways have detailed text descriptions that provide summaries of the pathway or the processes being depicted. Averaging a length of 342 words, these original descriptions highlight relevant background information, biological context and key reactions. Similarly, all compounds in PathBank have detailed text descriptions averaging 305 words. Further, all compounds have manually curated abbreviations for use in the easier-to-read large font pathway visualizations. Most of these compound names, abbreviations and descriptions were obtained from our own locally maintained databases such as HMDB (17), DrugBank (14), YMDB (18) and others. All protein names and descriptions are obtained from, and linked to, UniProt. Links and citation data for more than 12 958 different references (textbooks and journals) used to help create PathBank's pathways are also provided.

PathBank supports a wide variety of searching and browsing functions, including pathway browsing/searching (by name), compound browsing/searching (by name), protein browsing/searching (by name), sequence searching, compound similarity searching, molecular weight searching and advanced (Boolean) text searching. Users may also search PathBank by using lists of compound names, protein names and/or gene names or their corresponding identifiers (HMDB ID, Affymetrix ID, GenBank ID, UniProt ID, etc.). All of the pathways, proteins, and metabolites in PathBank are hyperlinked to other resources and all of the pathways are navigable through a modern visualization interface modeled after Google Maps (https://www.google.com/maps). PathBank also has a variety of tools to facilitate the display of specific pathway components and to map their relative concentrations. Similarly, different types of pathway displays are possible to facilitate printing or the preparation of slides and/or images. All of PathBank's pathway images, image maps, descriptions and reactions are freely downloadable in several common machine readable or data exchange formats including BioPAX, SBML, PWML, SBGN, RXN, PNG and SVG.

PathBank was implemented using a Ruby on Rails (http://rubyonrails.org, version 4.2.0) web framework incorporating a MySQL relational database (https://www.mysql.com, version 5.1.50) to manage all of the pathway data, including entity relationships, external references, descriptions, visualization specifications and chemical structures. PathBank uses the model–view–controller architecture, in which internal data logic is separated from user input and data presentation. The raw information stored in the database is dynamically extracted and rendered into web pages by PathBank's HTML interface responder. PathBank is hosted on a Digital Ocean server equipped with 4 CPUs, 50 GB of disk space, 8 GB of RAM and an additional 200 GB storage space on an Amazon S3 storage facility.

PathBank offers a number of features not found in other pathway databases. These features are highlighted in Table 1. In terms of the number of pathways per organism, PathBank offers the largest and possibly the most comprehensive collection of any public database. In terms of the total number of pathways, it is second only to KEGG (which has 653 007 pathways from 6076 organisms compared to the 10 organisms in PathBank). As shown in Table 1, PathBank also appears to provide much more visual detail, more biological context and more pathway-related data (text descriptions) than almost any other pathway database. In terms of biological context and detail, PathBank pathways illustrate organs, tissues, cells and organelles as well as physiological consequences (especially for drug action pathways). In terms of pathway-related data, PathBank provides almost twice as much text for its pathway descriptions as any other database, significantly more references than most other databases as well as uniquely providing both protein and metabolite descriptions (see Table 1). PathBank is particularly unique in its collection of lipid synthesis, disease, drug action, drug metabolism and metabolite signaling pathways. A key objective of PathBank is to provide pathway resources for the metabolomics and lipidomics communities. Technology now exists to detect and quantify thousands of individual lipids. Some of these lipids are unique to specific diseases, conditions, tissues or organisms. However, pathways describing the biosynthesis or degradation of these unique lipids are generally lacking in all current pathway databases (except PathBank). Indeed, because of these lipid pathways (as well as PathBank's many disease and drug action pathways) we estimate that >95% of the pathways found in PathBank are not found in any other pathway database. For instance, if pathway name similarity mapping is used to compare database overlap, <3% of PathBank's pathways are found in KEGG, <2% are found in Reactome, <1% are found in Wikipathways and <1% are found in BioCarta. Alternately, if one uses the average number of pathways/organism in each database and if one assumes all pathways in other databases are mostly covered in PathBank (a reasonable assumption) then only 1% of PathBank's pathways are found in KEGG, 11% are found in Reactome, 1% are found in Wikipathways and 3% are found in BioCarta.

PathBank's collection consists of a large number of lipid synthesis pathways. A major goal of PathBank is to have a pathway for every protein and a map for every metabolite. Lipids constitute more than 70% of the nameable metabolite entities in most organisms but unfortunately there is a tendency for many biochemists and pathway databases to ignore lipids or to treat them as generic molecules. From the perspective of classical metabolism, the use of generic classes is helpful. However, from the perspective of metabolomics and lipidomics, this approach is very limiting. Indeed, the primary data outputs of lipidomics and metabolomics experiments are individual compound identities and concentrations for hundreds to thousands of unique (but often similar) molecules. Converting these data into 'generics' or compound classes leads to a significant loss of information. Furthermore, recent work in lipidomics is highlighting how important individual lipid molecules are for certain biological functions or disease processes (19–21). This is why there has been a strong push within the metabolomics and lipidomics community to develop a database like PathBank where the goal is to have 'a pathway for every molecular entity'.

PathBank LAYOUT AND NAVIGATION

A screenshot montage of the PathBank interface and its search and browsing tools is shown in Figure 1. PathBank has a landing/home page very similar to that of its smaller cousin, SMPDB (15). This includes a short textual description of the database, a circulating 'carousel' of selected PathBank images and a series of tabs at the top of the page for navigation. An empty search box is located at the top left of each PathBank page. PathBank has five major tabs: (i)

Table 1. A comparison between PathBank and other pathway databases

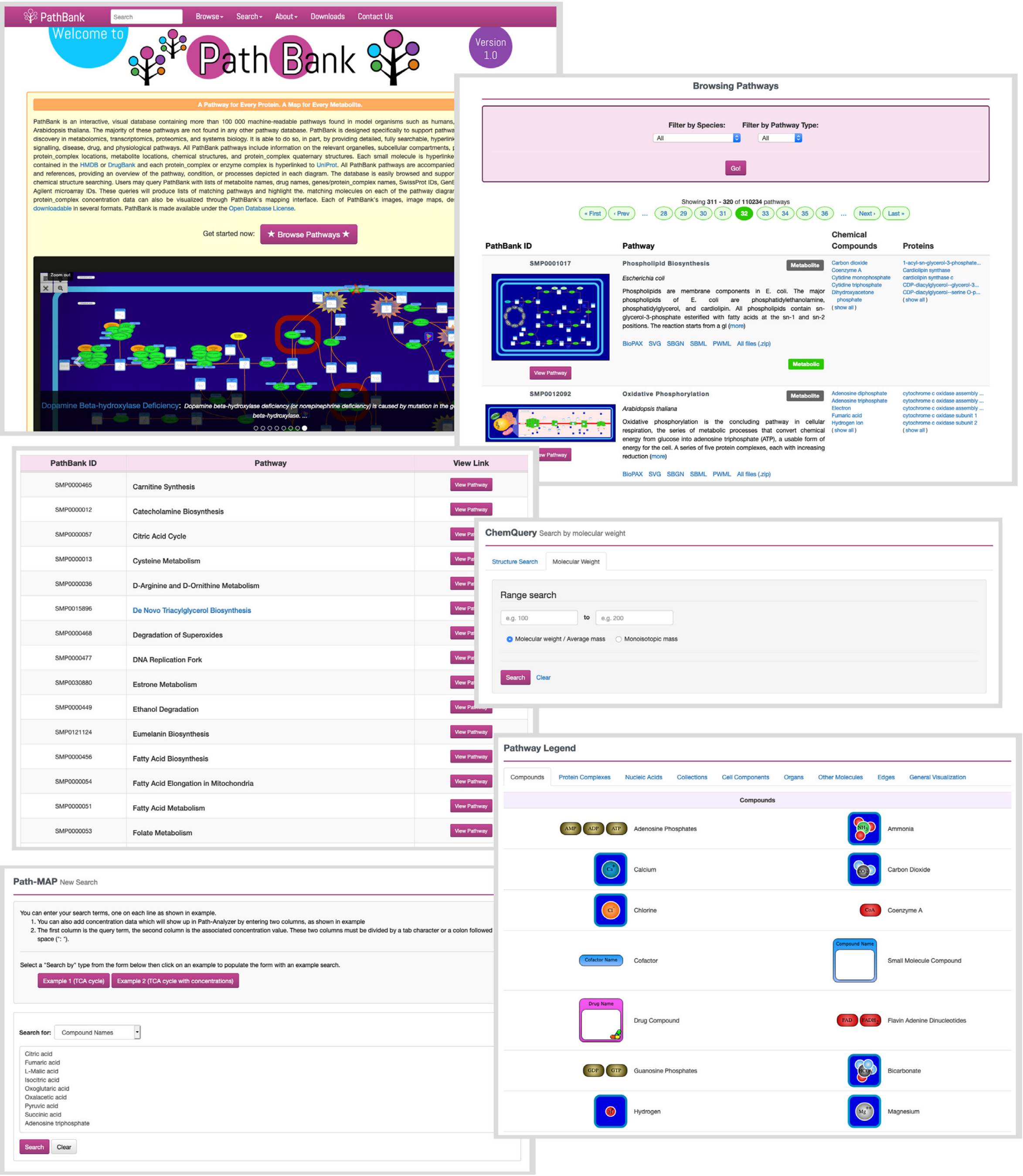
	PathBank	KEGG	Reactome	Wiki-Pathways	BioCarta
No. of organisms	10	6076	16	31	2
No. of pathways	110 234	653 007	20 882	2785	590
Ave. no. pathways/organism	1 1 023	107	1305	90	295
Protein pathways?	Yes	Yes	Yes	Yes	Yes
Metabolic pathways?	Yes	Yes	Yes	Yes	No
No. of pathways (Human)	48 701	335	2227	1,102	314
No. of pathways (*E. coli*)	1178	119	0	12	0
No. of pathways (*A. thaliana*)	3356	138	0	93	0
No. of pathways (Mouse)	12 651	331	1649	225	276
Hyperlinked proteins/genes	Yes	Yes	Yes	Yes	No
Hyperlinked metabolites	Yes	Yes	Yes	Yes	No
No. of disease pathways	20 692	91	411	118	5
No. of signaling pathways	64	65	718	211	39
No. of drug action pathways	404	75	46	39	0
Compound descriptions	Yes	No	No	No	No
Protein descriptions	Yes	No	No	No	No
Pathway summaries	Yes	Yes	Yes	Yes	No
Ave. word count per summary	342	68	179	143	0
Ave. ref. count per pathway	12	14	3	8	0
Rich visual display	Yes	No	Yes	No	Yes
Displays protein 4° structure	Yes	No	Yes	No	No
Displays metabolite structure	Yes	Yes	Yes	No	No
Displays cellular organelles	Yes	No	Yes	Yes	Yes
Displays cellular locations	Yes	No	Yes	Yes	Yes
Displays organs/tissues	Yes	No	Yes (in EHLDs)	No	No
Multicolored pathways	Yes	No	Yes	Yes	Yes
Multiple pathway views	Yes	No	No	No	No
No. pathway image formats	2	1	6	7	1
XML compatible	Yes	KGML	Yes	Yes	No
SBML compatible	Yes	No	Yes	Yes	No
BioPAX compatible	Yes	No	Yes	Yes	No
'Google Maps' interactivity	Yes	No	Yes	No	No
Community editable	Yes	No	No	Yes	No
Open access/Free download	Yes	No	Yes	Yes	Yes
Integrated with other DBs	Yes	Yes	Yes	Yes	No
Integrated with other Apps	Yes	Yes	Yes	Yes	No

Browse; (ii) Search; (iii) About; (iv) Downloads; and (v) Contact Us. Under the Browse tab users may browse the website using four different pages: a) Pathways; b) Table of Primary Pathways (TOPP); c) Compounds (metabolites or drugs); or d) Proteins. All pathway search and browse results are filterable by species (10 model species available) and by pathway type (Metabolite/Compound and Protein) using two dropdown filters. Additionally, selecting a pathway type will render visible a third dropdown menu to allow filtering of the results by pathway sub-categories (6 Metabolite/Compound pathway sub-categories and 15 Protein pathway sub-categories).

If users choose to browse by Pathway, a multi-page table is presented that contains four columns: (i) the PathBank identifier (with the thumbnail image); (ii) the pathway details, which includes the pathway name, species, description and download links; (iii) the compounds (which are hyperlinked); and (iv) the proteins (which are also hyperlinked). Clicking on the thumbnail image of a given pathway will direct users to the full-size pathway image. Likewise, clicking on the metabolite/compound names (column iii) or protein names (column iv) will open a page with detailed compound descriptions (from the appropriate organism-specific metabolome database such as HMDB or from a specialized organism-specific database) or protein descriptions (from UniProt).

If the TOPP option is selected, a multi-page table is presented showing three columns: (i) the PathBank identifier; (ii) the pathway name including the species name (if filtered by all species); and (iii) the link to the pathway view. In the TOPP, similarly named (replicated) pathways are grouped together to facilitate browsing of largely unique pathways. Groupings are represented by a single primary pathway, hence the table name, which has a clickable pathway name displayed in blue text. These hyperlinked pathways may be selected by the user to navigate to a popover view containing all of the related pathways which can be similarly browsed. In the case of lipid pathways, only the generic version of the pathway is listed in the main table; the pathways for the specific lipids are browsable in the generic lipid's popover view.

Users can browse PathBank's compounds and proteins in much the same manner. Multi-page tables present compounds and proteins alphabetically in three columns: (i) the compound/protein's identifier along with the corresponding View buttons; (ii) the compound/protein's name (with a short description); and (iii) the associated pathways (which are hyperlinked). Clicking on a compound's View button will take a user to the appropriate 'Compound Card', corresponding to the database most appropriate for the chosen (or filtered) organism. The default is HMDB if no organism has been selected. Clicking on a protein's corresponding View button will take a user to the appropriate 'UniProt



A Pathway for Every Protein. A Map for Every Metabolite.

PathBank is an interactive, visual database containing more than 100 000 machine-readable pathways found in model organisms such as humans, Arabidopsis thaliana. The majority of these pathways are not found in any other pathway database. PathBank is designed specifically to support pathwa discovery in metabolomics, transcriptomics, proteomics, and systems biology. It is able to do so, in part, by providing detailed, fully searchable, hyperlink signalling, disease, drug, and physiological pathways. All PathBank pathways include information on the relevant organelles, subcellular compartments, p protein_complex locations, metabolite locations, chemical structures, and protein_complex quaternary structures. Each small molecule is hyperlinke contained in the HMDB or DrugBank and each protein_complex or enzyme complex is hyperlinked to UniProt. All PathBank pathways are accompanied and references, providing an overview of the pathway, condition, or processes depicted in each diagram. The database is easily browsed and suppor chemical structure searching. Users may query PathBank with lists of metabolite names, drug names, genes/protein_complex names, SwissProt IDs, GenE Agilent microarray IDs. These queries will produce lists of matching pathways and highlight the. matching molecules on each of the pathway diagrar protein_complex concentration data can also be visualized through PathBank's mapping interface. Each of PathBank's images, image maps, de downloadable in several formats. PathBank is made available under the Open Database License.

Dopamine Beta-hydroxylase Deficiency: *Dopamine beta-hydroxylase deficiency (or norepinephrine deficiency) is caused by mutation in the g beta-hydroxylase. ...*

Showing **311** - **320** of **110234** pathways

PathBank ID	Pathway	Chemical Compounds	Proteins
SMP0001017	Phospholipid Biosynthesis *Escherichia coli* Phospholipids are membrane components in E. coli. The major phospholipids of E. coli are phosphatidylethanolamine, phosphatidylglycerol, and cardiolipin. All phospholipids contain sn-glycerol-3-phosphate esterified with fatty acids at the sn-1 and sn-2 positions. The reaction starts from a gl (more)	Carbon dioxide Coenzyme A Cytidine monophosphate Cytidine triphosphate Dihydroxyacetone phosphate (show all)	1-acyl-sn-glycerol-3-phosphate... Cardiolipin synthase cardiolipin synthase c CDP-diacylglycerol--glycerol-3... CDP-diacylglycerol--serine O-p... (show all)
SMP0012092	Oxidative Phosphorylation *Arabidopsis thaliana* Oxidative phosphorylation is the concluding pathway in cellular respiration, the series of metabolic processes that convert chemical energy from glucose into adenosine triphosphate (ATP), a usable form of energy for the cell. A series of five protein complexes, each with increasing reduction (more)	Adenosine diphosphate Adenosine triphosphate Electron Fumaric acid Hydrogen Ion (show all)	cytochrome c oxidase assembly ... cytochrome c oxidase assembly ... cytochrome c oxidase assembly ... cytochrome c oxidase subunit 1 cytochrome c oxidase subunit 2 (show all)

PathBank ID	Pathway	View Link
SMP0000465	Carnitine Synthesis	View Pathway
SMP0000012	Catecholamine Biosynthesis	View Pathway
SMP0000057	Citric Acid Cycle	View Pa
SMP0000013	Cysteine Metabolism	View Pa
SMP0000036	D-Arginine and D-Ornithine Metabolism	View Pa
SMP0015896	De Novo Triacylglycerol Biosynthesis	View Pa
SMP0000468	Degradation of Superoxides	View Pa
SMP0000477	DNA Replication Fork	View Pa
SMP0030880	Estrone Metabolism	View Pa
SMP0000449	Ethanol Degradation	View Pathway
SMP0121124	Eumelanin Biosynthesis	View Pathway
SMP0000456	Fatty Acid Biosynthesis	View Pathway
SMP0000054	Fatty Acid Elongation in Mitochondria	View Pathway
SMP0000051	Fatty Acid Metabolism	View Pathway
SMP0000053	Folate Metabolism	View Pathway

ChemQuery Search by molecular weight

Range search

e.g. 100 to e.g. 200

● Molecular weight / Average mass ○ Monoisotopic mass

Pathway Legend — Compounds

AMP ADP ATP	Adenosine Phosphates	NH_3	Ammonia
Ca^+	Calcium	CO_2	Carbon Dioxide
Cl^-	Chlorine	CoA	Coenzyme A
Cofactor Name	Cofactor	Compound Name	Small Molecule Compound
Drug Name	Drug Compound	FAD $FADH_2$	Flavin Adenine Dinucleotides
GDP GTP	Guanosine Phosphates	HCO_3^-	Bicarbonate
H^+	Hydrogen	Mg^{++}	Magnesium

Path-MAP New Search

You can enter your search terms, one on each line as shown in example.

1. You can also add concentration data which will show up in Path-Analyzer by entering two columns, as shown in example
2. The first column is the query term, the second column is the associated concentration value. These two columns must be divided by a tab character or a colon followed space (": ").

Select a "Search by" type from the form below then click on an example to populate the form with an example search.

Search for: Compound Names

Citric acid
Fumaric acid
L-Malic acid
Isocitric acid
Oxoglutaric acid
Oxalacetic acid
Pyruvic acid
Succinic acid
Adenosine triphosphate

Figure 1. A screenshot montage of different browsing and searching screens taken from PathBank. A more detailed description of the different functions and capabilities of the various browse and search tools in PathBank is given in the text.

Card'. Unlike the pathway browse results, compounds and proteins are filterable only by species using the dropdown menu provided. An additional filter in the form of a Search box at the top right of the table allows users to filter the page by compound, protein or pathway name.

PathBank places a strong emphasis on providing users with high quality, artistically pleasing pathway diagrams that are not only correct and informative but also colorful, interactive and richly detailed. All of the pathway diagrams in PathBank use scalable vector graphics (SVG) and a web interface technology inspired by Google Maps. This allows rapid and continuous zooming using a mouse pad or a mouse scroll wheel or through simply clicking on-screen zoom icons. It also allows facile navigation around zoomed-in pathway diagrams through a simple click-and-drag operation or through clicking on-screen up/down or left/right arrows located near the on-screen zoom functions. A full-screen view of each PathBank pathway diagram is also available, which can be toggled off and on by clicking the full-screen icon located between the navigation arrows.

On the right side of each pathway diagram is a pathway display panel with five tabs (Description, Highlight, Analyze, Downloads, Settings/Display). The default view is the Description panel, which describes the pathway in detail and provides one or more literature references. The Highlight panel (viewed by clicking the 'Highlight' tab) allows users to select and color different proteins and metabolites for display purposes. The Analyze panel (viewed by clicking the 'Analyze' tab) allows users to enter concentration (relative or absolute) data on proteins/transcripts and/or metabolites and to have these colored on the pathway diagram. It is highly recommended that users switch to the black and white version of the pathway when using the highlight and analyze tools in order to clearly see the elements being colored. The download panel (viewable by clicking the 'Download' tab) allows users to select the type of file format that they wish to have their pathway image (annotated or untouched) saved as. Options are available for downloading images in BioPAX (22) image format (full color SVG, grayscale SVG, simple SVG, large font SVG, and simple large font SVG), BioPAX only format, SBGN (23), SBML (24), PWML (a custom PathWhiz format), PNG or all of the above. At last, the Display panel (viewed by clicking the 'Display' tab) allows users to change the pathway display to suit a particular taste or application. Users may toggle between simple (thick line) and complex (individual lipid) membranes; between a dark blue (aqueous) background or a white background; between the full color, highly detailed version and the simplified KEGG-like (black and white) representation; and finally between full-names of pathway components and their large font abbreviations. There is also an additional fifth option to render the simplified (KEGG-like) pathway representation using the easier-to-read large font abbreviations.

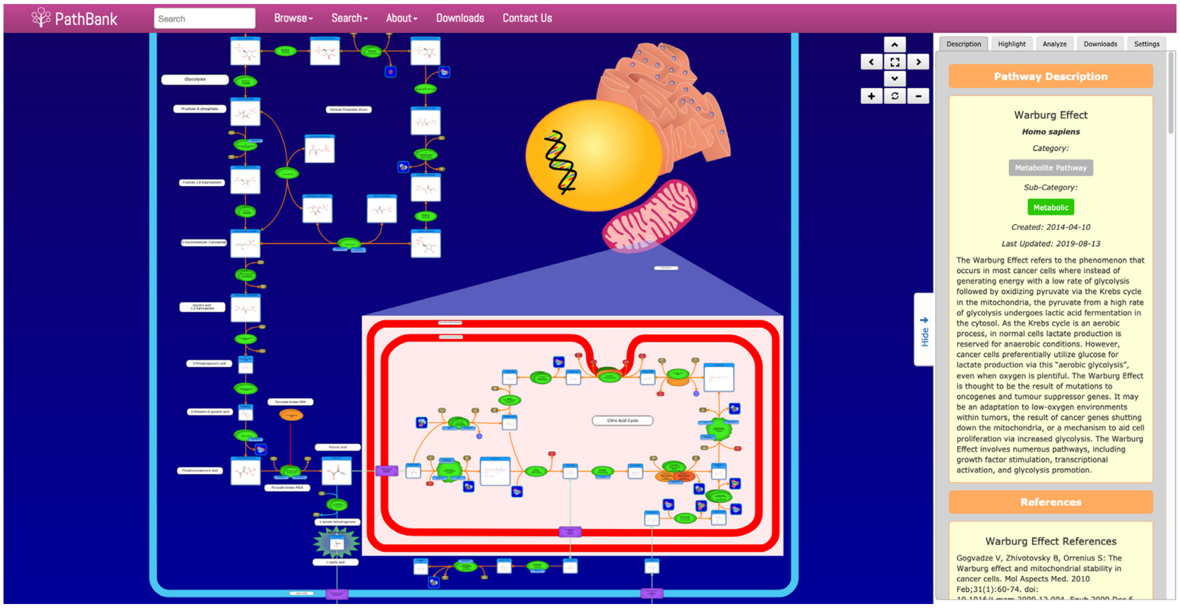
In addition to these browsing functions, PathBank also offers extensive search functions. These include a general text search (available at the top of every PathBank web page) as well as several, more specific, search functions listed under the 'Search' tab. The text search supports Boolean logic (AND, OR and NOT operations) along with field-specific searches covering compound names, protein names, compound/protein identifiers, pathway names, and pathway descriptions. Instructions for the text search are accessible using the Search dropdown list. There are four additional searches available: (i) the Path-MAP Advanced Search; (ii) the ChemQuery Structure Search; (iii) the Molecular Weight Search; and (iv) the Sequence Search. The Path-MAP search allows users to enter long lists of compound names, protein names, compound identifiers, protein identifiers or gene identifiers to search for organism-specific pathways enriched with these entities. The result is a list of pathways with enrichment scores calculated using the frequency of matches and the number of pathway entities as scaled using a hypergeometric function. It is also possible to enter similar lists with Path-MAP but with concentrations or relative concentrations, as might be obtained from a typical proteomics, transcriptomics or metabolomics experiment. The result of this type of search is a pathway annotated and colored with the corresponding metabolite/protein concentrations according to a yellow (low)-red (high) concentration gradient. Other searches such as the ChemQuery and Molecular Weight searches are intended for metabolite or compound searching against PathBank's chemical database of 78 271 compounds and are identical to those used by HMDB, SMPDB and many other databases. The Sequence search uses BLAST (25) to find sequence matches or sequence similarities to the protein sequences within PathBank's sequence database of 8973 proteins.

PathBank's 'About' tab provides additional information about PathBank, its associated release notes, the required citations, database statistics, the PathBank style guide, links to other Pathway Databases, as well as images (via a Pathway legend) for the different components (organelles, organs, tissues) seen in PathBank pathways. PathBank's 'Download' tab allows users to download pathways, metabolite names and protein names (in CSV files) as well as all of its pathways in BioPAX, SBGN, SBML and PWML format. A subset of PathBank's pathways (i.e. the primary pathways browsable through the TOPP) are also downloadable in SVG and PNG format. Due to the extremely large file size, users who wish to download all pathway images will need to submit a request using PathBank's Contact form. All of PathBank's sequences (both gene and protein) are available in FASTA text format, all of its chemical structures are available in SDF format and all of its reactions are available in RXN format.

PathBank ASSEMBLY, QUALITY CONTROL AND CURATION

PathBank was assembled using the PathWhiz pathway illustrator (16) using a team of pathway curators who have been manually researching, illustrating and refining pathway diagrams for more than 5 years. PathBank was initially built from a number of small molecule (metabolite) pathway diagrams that were originally assembled for SMPDB (15), HMDB (17), DrugBank (14), YMDB (18) and ECMDB (26). These pathway databases are restricted to depicting the metabolic or metabolite signaling pathways associated with humans (SMPDB, HMDB, DrugBank), yeast (YMDB) and *E. coli* (ECMDB). PathBank extends these existing pathway collections to cover other model organisms and to include disease, physiological, protein and cellular signaling pathways. The PathBank curation team took advantage of a number of recent improvements to the PathWhiz pathway illustrator (16,27). In particular, the latest version of PathWhiz has been enhanced to permit near-automated pathway propagation (to create homologous pathways for other species) and pathway replication (to create similar pathways within a species). PathWhiz has also been modified to support protein signaling, protein–protein interactions, protein–DNA interactions, transport and transcriptional events. An example of a PathBank pathway diagram generated via PathWhiz (illustrating the Warburg effect) is shown in Figure 2.

As described previously (15,16), all pathways constructed via PathWhiz are built according to standard operating protocols (SOPs), checklists and pathway illustration style guides to ensure general uniformity and consistency across pathways and across species. The comprehensive PathBank style guide, complete with numerous visual examples, is



Pathway Description

Warburg Effect

Homo sapiens

Category:

Metabolite Pathway

Sub-Category:

Metabolic

Created: 2014-04-10

Last Updated: 2019-08-13

The Warburg Effect refers to the phenomenon that occurs in most cancer cells where instead of generating energy with a low rate of glycolysis followed by oxidizing pyruvate via the Krebs cycle in the mitochondria, the pyruvate from a high rate of glycolysis undergoes lactic acid fermentation in the cytosol. As the Krebs cycle is an aerobic process, in normal cells lactate production is reserved for anaerobic conditions. However, cancer cells preferentially utilize glucose for lactate production via this "aerobic glycolysis", even when oxygen is plentiful. The Warburg Effect is thought to be the result of mutations to oncogenes and tumour suppressor genes. It may be an adaptation to low-oxygen environments within tumors, the result of cancer genes shutting down the mitochondria, or a mechanism to aid cell proliferation via increased glycolysis. The Warburg Effect involves numerous pathways, including growth factor stimulation, transcriptional activation, and glycolysis promotion.

References

Warburg Effect References

Gogvadze V, Zhivotovsky B, Orrenius S: The Warburg effect and mitochondrial stability in cancer cells. Mol Aspects Med. 2010 Feb;31(1):60-74. doi: 10.1016/j.mam.2009.12.004. Epub 2009 Dec 6

Figure 2. A screenshot from PathBank's 'Warburg Effect' pathway. First identified in the 1920's, this is a pathway commonly activated in cancer and it is classified as a 'Metabolite Disease' pathway. Curiously, no other commonly used pathway database appears to illustrate this very well-known and important pathway.

posted on the web page under the 'About' tab. In selecting which pathways would be drawn for a given species, members of the curation team would first perform an extensive literature search to identify unique metabolite classes or metabolic processes for the organism of interest. Second, a literature review would be done to search for unique proteins or unique protein/cellular signaling processes. Next, a detailed review of existing pathway diagrams (both metabolic and protein signaling) from multiple public pathway databases for that organism would be performed. These results would then be compiled and compared against existing internal reference pathway databases corresponding to the closest species (such as the HMDB for other mammals) to that organism. Further discussions and team meetings would be held wherein curators would discuss their findings with PathBank's lead curator (CL) and its principal investigator (DSW). Through this process new pathways for a given organism would be identified, selected (or rejected), assessed for their similarity to existing pathways, sketched (by hand) and finally manually drawn via PathWhiz (if unique) or replicated via PathWhiz and modified (if similar to another species) by members of the PathBank curation team.

After a given pathway was illustrated using PathWhiz, it would be evaluated independently by a second pathway curator (often the lead curator or the PI) to ensure it adhered to the PathWhiz style guides and SOPs. If the pathway failed to meet style guide requirements or failed to follow the SOPs, it would be corrected and re-evaluated until it met those requirements. Only after a pathway diagram passed secondary inspection would it be allowed to be propagated (to other species) or replicated (for related pathways within a species). This quality control protocol was applied to all hand-drawn primary pathways in PathBank which serve as template pathways that can be automatically replicated and/or propagated via PathWhiz. Spot checks (of ∼5% of the pathways) for any automatically replicated/propagated pathways were manually done by the lead curator(s) to ensure logical and biological consistency. Global or system-wide checks were also done periodically to ensure that organelle or organ/tissue images were appropriate for a given organism. Many of the primary pathways (especially for metabolism) that have been online for several years have undergone heavy remediations in order to comply with the guidelines outlined in PathBank's detailed style guide. As a result of this huge effort, these redrawn pathways have a standard, more aesthetically pleasing appearance, greater biological context (e.g. transporters across cell and organelle membranes) and increased connectivity between pathway entities. Many pathways have been updated with new descriptions and icons that have been drawn for PathBank's growing image collection. User feedback has also been used to further refine and improve these pathways.

All members of the PathBank curation team were required to have at least an undergraduate degree in bioinformatics or molecular biology. This ensured that they had sufficient biological and/or biochemical knowledge to understand the scientific literature and the biological processes being illustrated. All curation team members were also given extensive training by the lead curator(s) in pathway illustration via hands-on mentoring, text instructions, peer support, and video tutorials. However, it is important to note that given the inherently artistic nature of pathway illustration, the differences in artistic/illustrative skills among the curation team and the complexity of some biological processes, it is impossible to guarantee that all pathways in PathBank exhibit the same level of visual or conceptual quality.

Improvements and updates to PathBank's pathways are an ongoing process. Minor corrections or small improvements in image or pathway quality or pathway descriptions will be done without a formal update announcement. However, significant changes, additions or improvements to an individual pathway diagram (>10% in the number of pathway components), will be listed in the pathway description and the last updated date will be modified to reflect any such changes. As this is only version 1.0 of the PathBank database, all pathway diagrams are dated with August 2019 as the last update date. Large-scale updates and improvements to the database in the future will be given database version numbers (2.0, 3.0, etc.) and suitable database update dates. They will also be described in detail as publications or online update descriptions as appropriate.

PathBank is made available under the Open Database License in which users are free to contribute to the database using its PathWhiz pathway drawing tool. Indeed, one of the authors listed on this manuscript (MRG) contributed >50 primary pathways to PathBank as an external 'community' member. Of course, not all community contributions are of sufficient quality to merit inclusion to PathBank. Any decision to include/exclude community or crowd-sourced pathway submissions depends on the quality/quantity of the contribution and their adherence to PathBank's SOPs and style guide. The final decision for any external contribution is ultimately made by PathBank's lead curator. The PathBank curation team will closely work with, and will train, external pathway illustrators who are serious about contributing to this resource.

PathBank LIMITATIONS AND FUTURE PLANS

Like all pathway databases, PathBank is a work in progress. Currently both the quantity and level of coverage of PathBank's protein pathways falls far short of what is currently available for PathBank's metabolite/compound pathways. Indeed, only 398 protein pathways have been illustrated to date, compared to 109 836 metabolite pathways. This imbalance reflects the past focus of the PathBank team's curation effort (which was primarily in metabolism), the extremely large number of lipid pathways (>90 000) and the challenges that exist in depicting the somewhat more complex protein signaling pathways. As PathBank's collection expands, so too will its genome coverage (currently ranging between 1.9 and 25.1%) for each model organism (viewable on PathBank's statistics page). Of course, even with this major annotation effort, not every known pathway will appear in PathBank—now or in the near future. Likewise, it is expected that a number of users will notice that some PathBank pathways do not depict all expected or relevant entities in the pathway diagrams. In these situations, we certainly encourage users to contact the PathBank curation team (via the 'Contact' tab) so that those corrections can be made or those PathBank depictions can be more fully explained.

Another clear limitation with PathBank is the relatively small number of model organisms that are currently covered. Currently, only 10 organisms are included in PathBank's 'zoo' with several well-known model organisms, including a few key vertebrates, several parasites and a number of important bacteria, not being included. Over the coming two years, it is expected that at least nine other model organisms will be added to the database, including the zebrafish (*Danio rerio*), the African clawed frog (*Xenopus laevis*), the chicken (*Gallus gallus*), rice (*Oryza sativa*), the malaria parasite (*Plasmodium falciparum*), the sleeping sickness parasite (*Trypanosoma brucei*) and several bacteria, including *Bacillus subtilis*, *Staphylococcus aureus*, and *Thermotoga maritime*. Additional organisms or species may be added depending on user feedback.

PathBank is also somewhat limited in the analytical tools that it offers. This limited offering was a deliberate decision. Because all of PathBank's data are freely downloadable, we believe it would be far better for PathBank to serve as a data resource hub rather than an analytical hub. By making PathBank more of a data resource hub, we hope to encourage bioinformatics specialists to download or use the PathBank data to develop and implement their own, much more sophisticated analytical tools and web servers. Indeed, this has already happened. For instance, PathBank's data is now being used to support a very popular analytical web server called MetaboAnalyst 4.0 (28).

PathBank was originally designed to facilitate interactive, web-based visualization and pathway navigation. However, this has led to the creation of pathways that are not particularly 'printer friendly'. Indeed, many PathBank pathways have entity labels that are generally too small to be seen in conventional slides or manuscript figures. To improve the situation, PathBank now offers two alternative pathway displays (available through the Settings tab of the Display panel) that allows users to toggle between 'regular text' and 'large text' for both the full-color and simplified visualizations so that the image labels are more visible and printer-friendly. In order to accommodate the larger font size, manually curated abbreviations and gene symbols adhering to a strict seven-character limit are used in place of the longer full names of compounds and proteins. Additional efforts are ongoing to create even smaller or more compact pathway visualizations with less empty space and more grid-like entity layouts.

CONCLUSION

PathBank represents a new type of pathway database. It has been designed to take some of the best features of earlier pathway databases and to combine them with newer, more modern kinds of web visualization tools and more advanced web interfaces to create a fully open access, visually rich and highly comprehensive pathway resource. PathBank also brings a number of unique features into the Pathway database field including exceptional pathway coverage, richly detailed pathway and entity descriptions, a number of new pathway types or categories (metabolite signaling, drug action, drug metabolism, disease-specific pathways), new kinds of pathway or entity visualization details (cellular location, organelles, organs, protein quaternary structure) and a much wider variety of pathway viewing and data exchange/storage options. Being the first release of this particular pathway database, we expect that users will note some missing elements or inconsistences with certain pathways or pathway data. However, we anticipate that through user feedback, continued community engagement and ongoing curation efforts, the next release of PathBank will be even more comprehensive and potentially much more useful to the community.

ACKNOWLEDGEMENTS

We would like to thank Aidin Foroutan and Noah A. Wishart for their contributions to the earliest versions of PathBank.

FUNDING

Canadian Institutes for Health Research (CIHR); Canada Foundation for Innovation, Major Science Initiatives Fund (CFI-MSI); Genome Alberta (a division of Genome Canada). Funding for open access charges: Genome Alberta.

Conflict of interest statement. None declared.

REFERENCES

1. Sundararaj,S., Guo,A., Habibi-Nazhad,B., Rouani,M., Stothard,P., Ellison,M. and Wishart,D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. *Nucleic Acids Res.*, **32**, D293–D295.
2. Orth,J.D., Thiele,I. and Palsson,B.Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.*, **28**, 245–248.
3. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
4. Xia,J. and Wishart,D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.*, **38**, W71–W77.
5. Leong,H.S. and Kipling,D. (2009) Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Res.*, **37**, e79.
6. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
7. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
8. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
9. Kelder,T., van Iersel,M.P., Hanspers,K., Kutmon,M., Conklin,B.R., Evelo,C.T. and Pico,A.R. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.
10. Nishimura,D. (2001) A view from the web: Biocarta. *Biotech. Softw. Internet Rep.*, **2**, 117–120.
11. Breuer,K., Foroushani,A.K., Laird,M.R., Chen,C., Sribnaia,A., Lo,R., Winsor,G.L., Hancock,R.E., Brinkman,F.S. and Lynn,D.J. (2013) InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res.*, **41**, D1228–D1233.
12. Barbarino,J.M., Whirl-Carrillo,M., Altman,R.B. and Klein,T.E. (2018) PharmGKB: a worldwide resource for pharmacogenomic information. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **10**, e1417.
13. 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.
14. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
15. Jewison,T., Su,Y., Disfany,F.M., Liang,Y., Knox,C., Maciejewski,A., Poelzer,J., Huynh,J., Zhou,Y., Arndt,D. *et al.* (2014) SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Res.*, **42**, D478–D484.
16. Pon,A., Jewison,T., Su,Y., Liang,Y., Knox,C., Maciejewski,A., Wilson,M. and Wishart,D.S. (2015) Pathways with PathWhiz. *Nucleic Acids Res.*, **43**, W552–W559.
17. Wishart,D.S., Feunang,Y.D., Marcu,A., Guo,A.C., Liang,K., Vázquez-Fresno,R., Sajed,T., Johnson,D., Li,C., Karu,N. *et al.* (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.
18. Ramirez-Gaona,M., Marcu,A., Pon,A., Guo,A.C., Sajed,T., Wishart,N.A., Karu,N., Djoumbou Feunang,Y., Arndt,D. and Wishart,D.S. (2017) YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res.*, **45**, D440–D445.
19. Proitsi,P., Kim,M., Whiley,L., Simmons,A., Sattlecker,M., Velayudhan,L., Lupton,M.K., Soininen,H., Kloszewska,I., Mecocci,P. *et al.* (2017). Association of blood lipids with Alzheimer's disease: a comprehensive lipidomics analysis. *Alzheimer Dement.*, **13**, 140–151.
20. Zhao,Y., Cheng,X. and Lin,R. (2014). Lipidomics applications for discovering biomarkers of diseases in clinical chemistry. *Int. Rev. Cell Mol. Biol.*, **313**, 1–26.
21. Afshinnia,F., Rajendiran,T., Wernisch,S., Soni,T., Jadoon,A., Karnovsky,A., Michailidis,G. and Pennathur,S. (2018). Lipidomics and biomarker discovery in kidney disease. *Semin. Nephrol.*, **38**, 127–141.
22. 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. Biotech.*, **28**, 935–942.
23. Le Novere,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. Biotech.*, **27**, 735–741.
24. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
25. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
26. Sajed,T., Marcu,A., Ramirez,M., Pon,A., Guo,A.C., Knox,C., Wilson,M., Grant,J.R., Djoumbou,Y. and Wishart,D.S. (2016) ECMDB 2.0: a richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Res.*, **44**, D495–D501.
27. Ramirez-Gaona,M., Marcu,A., Pon,A., Grant,J., Wu,A. and Wishart,D.S. (2017) A web tool for generating high quality machine-readable biological pathways. *J. Vis. Exp.*, **8**, 120.
28. Chong,J., Soufan,O., Li,C., Caraus,I., Li,S., Bourque,G., Wishart,D.S. and Xia,J. (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.*, **46**, W486–W494.