Nucleic Acids Research, 2023, 1–9
https://doi.org/10.1093/nar/gkad1041
Database issue

OXFORD

PathBank 2.0—the pathway database for model organism metabolomics

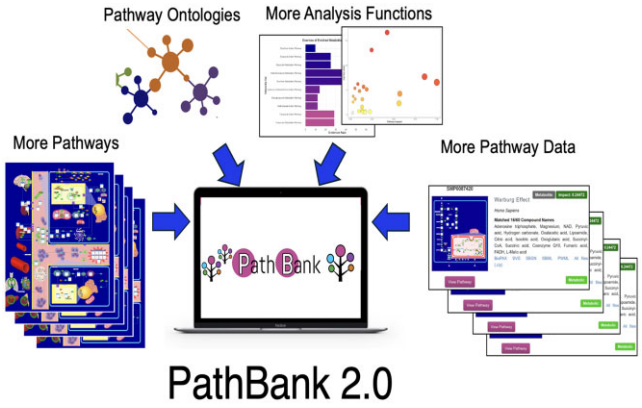
David S. Wishart[1,2,3,4,*], Ray Kruger[1], Aadhavya Sivakumaran[1], Karxena Harford[1], Selena Sanford[1], Rahil Doshi[1], Nitya Kehrtarpal[1], Omolola Fatokun[1], Daphnee Doucet[1], Ashley Zubkowski[1], Hayley Jackson[1], Gina Sykes[1], Miguel Ramirez-Gaona[5], Ana Marcu[1], Carin Li[1], Kristen Yee[1], Christiana Garros[1], Dorsa Yahya Rayat[1], Jeanne Coleongco[1], Tharuni Nandyala[1], Vasuk Gautam[1] and Eponine Oler[1]

[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
[3]Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, AB T6G 2B7, Canada
[4]Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB T6G 2H7, Canada
[5]Department of Plant Breeding, Wageningen University and Research, 6708 PB Wageningen, Gelderland, Netherlands
[*]To whom correspondence should be addressed. Tel: +1 780 492 8574; Email: dwishart@ualberta.ca

Abstract

PathBank (https://pathbank.org) and its predecessor database, the Small Molecule Pathway Database (SMPDB), have been providing comprehensive metabolite pathway information for the metabolomics community since 2010. Over the past 14 years, these pathway databases have grown and evolved significantly to meet the needs of the metabolomics community and respond to continuing changes in computing technology. This year's update, PathBank 2.0, brings a number of important improvements and upgrades that should make the database more useful and more appealing to a larger cross-section of users. In particular, these improvements include: (i) a significant increase in the number of primary or canonical pathways (from 1720 to 6951); (ii) a massive increase in the total number of pathways (from 110 234 to 605 359); (iii) significant improvements to the quality of pathway diagrams and pathway descriptions; (iv) a strong emphasis on drug metabolism and drug mechanism pathways; (v) making most pathway images more slide-compatible and manuscript-compatible; (vi) adding tools to support better pathway filtering and selecting through a more complete pathway taxonomy; (vii) adding pathway analysis tools for visualizing and calculating pathway enrichment. Many other minor improvements and updates to the content, the interface and general performance of the PathBank website have also been made. Overall, we believe these upgrades and updates should greatly enhance PathBank's ease of use and its potential applications for interpreting metabolomics data.

Graphical abstract



Introduction

Pathway analysis is becoming increasingly important for the interpretation of metabolomics, proteomics and transcriptomics data (1–4). In particular, pathway analysis provides not only important biological and spatial context to molecular omics measurements, but it allows much more facile integration of metabolite, protein and genetic data. Key to the success of any pathway analysis is the availability of

Received: September 15, 2023. **Revised:** October 19, 2023. **Editorial Decision:** October 20, 2023. **Accepted:** October 31, 2023

high-quality pathway databases containing a broad diversity of pathways and richly illustrated pathway diagrams. The ideal pathway database should be comprehensive, covering most known pathways, most known metabolites (and xenobiotics) as well as most known genes or proteins for a given organism. The pathways in such a database should cover a broad number of categories including metabolic (anabolic/catabolic), molecular signaling, endocrine, physiological, drug action, drug metabolism and disease mechanism pathways. Ideally the pathways in these databases should contain information and images indicating the spatial locations of these pathways including the relevant organs, tissues, cells, cell compartments and organelles. Similarly, structural information about the relevant chemicals or proteins involved in the pathways and descriptions about the pathways themselves should be available. Preferably, the pathways should be both human readable and machine readable, freely available in many different formats, openly downloadable and amenable for slide presentations or figures in papers.

PathBank was developed to be just such an 'ideal' pathway database. Specifically, PathBank is a comprehensive, richly illustrated, deeply annotated, machine-readable pathway database that was specifically developed for the interpretation of metabolomics, proteomics and transcriptomics data. It was first described in 2020 with the release of version 1.0 (5). PathBank actually evolved from an earlier human-specific metabolic pathway database called the Small Molecule Pathway Database or SMPDB (6,7). All the pathways contained in PathBank and SMPDB have been painstakingly designed and carefully hand-illustrated by a highly trained team of pathway curators using strict protocols and a specially developed, open access, web-based, pathway illustration tool called PathWhiz (8,9). PathWhiz uses a large palette of pre-generated organ, tissue, cellular, subcellular and molecular images, as well as various automated sequence/structure search and annotation tools to enable the rapid, visually consistent rendering of molecular pathways that are both human and machine readable. PathWhiz also allows pathways to be easily replicated (within species) or propagated (across species) to save time and reduce unnecessary repetition when rendering PathBank pathways.

PathBank was primarily developed to address the limitations and shortcomings of other well-known pathway databases such as KEGG (10,11), the Cyc databases (12–14), Reactome (15) and Wikipathways (16,17) – specifically with respect to metabolomics applications and metabolite coverage. Indeed, prior to the introduction of SMPDB or PathBank, no pathway database had been developed to address the specific needs of the metabolomics community. However, in the course of addressing the specific needs of the metabolomics community, we found that PathBank also helped address many of the unmet needs of the proteomics and transcriptomics/genomics communities as well. The first release of PathBank contained >110 000 metabolic pathways for 10 model organisms (*H. sapiens, B. taurus, R. norvegicus, M. musculus, D. melanogaster, C. elegans, A. thaliana, S. cerevisiae*, *E. coli* and *P. aeruginosa*). These pathways described the interactions of nearly 80 000 different compounds (metabolites, drugs and other xenobiotics), 9000 different proteins and genes, and >175 000 different reactions and interactions. Some of the more unique features of PathBank (version 1.0) that made it particularly appealing to the metabolomics (and other omics) communities were its level of metabolite comprehensiveness (5–10× more metabolite coverage than other databases), its broad scope of pathway types (catabolism/anabolism; physiological/endocrine; metabolite signaling; drug action; drug metabolism and metabolic disease), its wide variety of searching and browsing functions, the rich cellular, subcellular and physiological information contained within its pathways, its detailed pathway and pathway component descriptions and its compatibility with several common machine readable or data exchange formats including BioPAX (18), SBML (19), PWML (8,9), SBGN (20), PNG and SVG. Additional details regarding PathBank's (version 1.0) distinguishing features with regard to other well-known pathway databases (11,14,15,17,21–24), and additional information about its exact content, uniqueness, target audience, use cases, curation methods, layout and design are provided in the original PathBank paper (5).

Since its release in 2020, PathBank has proven to be particularly popular. Its pathways and pathway information have been integrated into recent releases of a number of very popular omics databases such as the HMDB (25), DrugBank (26), MarkerDB (27), DAVID (28) and MetaboAnalyst (29–33). Likewise, PathBank and SMPDB continue to be very well cited and heavily used by a wide range of users from around the world. However, as with any large-scale, online biological database, there is always room for improvement and always a need to expand, upgrade or improve the data content. Version 2.0 of PathBank has added over 5000 primary (also called canonical) pathways, and nearly 600 000 primary and secondary (i.e. total) pathways. We have also received some excellent feedback from the user community and the developers of the secondary databases and software tools that use PathBank as part of their data analysis pipelines or database search tools.

Among the areas identified as needing expansion or improvement in PathBank were: (i) the quality and completeness of pathway diagrams and descriptions; (ii) the coverage of drug metabolism and drug mechanism pathways; (iii) the breadth of physiological and signaling pathways; (iv) the compatibility of pathways for slide and manuscript presentations and (v) the addition of more tools for filtering, navigating, selecting and analyzing pathways. Additionally, our own team identified much needed hardware and software upgrades as well as a number of 'back-end' innovations that improved the overall performance of the database. As a result, we have done our best to implement these suggested/required improvements into the latest release of PathBank, known as PathBank 2.0. In the following pages, we describe the updates and upgrades to PathBank 2.0 under the following four subsections: (i) improvements in layout, implementation and curation; (ii) pathway additions; (iii) pathway improvements and (iv) new pathway analysis tools.

Improvements in layout, implementation and curation

Improved layout

The PathBank 2.0 layout has been modified to improve navigation and accelerate entity selection. In particular, the new PathBank 2.0 landing page features an updated database description, a rotating image carousel, and a Navigation Bar

at the top of the page with six key tabs: 'Browse', 'Search', 'Analyze', 'About', 'Downloads' and 'Contact Us'. An empty search box is situated at the top left of the Navigation Bar. By clicking the 'Browse' tab, users can explore four different options: Pathways, the Table of Primary Pathways (TOPP), Compounds (metabolites or drugs), or Proteins. Recall that in PathBank 2.0, pathways are separated into primary or canonical pathways (listed in TOPP) and secondary or daughter pathways (listed in Pathways). *Primary Pathways* are those pathways that describe a distinct, broadly thematic biological process, such as protein synthesis. In PathBank 2.0, there are now 6951 of these kinds of pathways. *Secondary Pathways* are those that are derived from primary pathways and contain only minor modifications to the primary pathway, such as protein synthesis with phenylalanine incorporation. In PathBank 2.0 there are now >590 000 of these 'regular' or secondary pathways. Each pathway in PathBank 2.0 is associated with one or more species and each species-specific pathway contains either proteins alone, metabolites/chemicals alone or both metabolites and proteins together. On average, each pathway in PathBank contains 21 metabolites and 7 enzymes or proteins.

In PathBank 2.0, all search and browsing results can be further filtered by species and pathway type through different dropdown menus located at the top of each 'Browse' or 'Search' page. Under the pathway type filter there are three super-categories (all, metabolite and protein), six metabolite pathway subcategories and 15 protein pathway subcategories. The six metabolite subcategories are: (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). The 15 protein subcategories are: (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.

For users opting to browse **Pathways**, a multi-page table with five columns is presented that lists the: (i) *PathBank ID* (with a thumbnail image); (ii) *Pathway Name and Description* (with hyperlinks to machine-readable pathway files); (iii) *Pathway Class*; (iv) associated *Chemical Compounds* and (v) associated *Proteins*. The table is sortable by *PathBank ID, Pathway Name* and *Pathway Class* by clicking on the up/down arrows in the column titles. The *Pathway Class* is a new taxonomic pathway grouping for PathBank 2.0 and represents a more detailed pathway classification layer below the super-category and sub-category. There are a total of 109 different *Pathway Classes* in PathBank 2.0. The full pathway taxonomy for PathBank 2.0 is shown under the 'About' tab (see PathBank Taxonomy). Definitions for the super-categories, sub-categories and classes have been added under the 'About' tab (see Pathway Category Definitions), as well as mappings to applicable NCBO Pathway Ontology classes (34). Navigation through the content of the Pathway Table is simple (via hyperlinks) and self-explanatory. Clicking on each pathway thumbnail image opens up the full-scale, fully navigable pathway diagram in a new window. Each pathway diagram has a display panel with tabs titled 'Description', 'Highlight', 'Analyze', 'Downloads' and 'Settings/Display'. Users can read more about the pathway (Description), select and color proteins or metabolites for display (Highlight), input concentration data for coloring (Analyze), and customize the pathway display (Settings/Display). Significant improvements have been made to the pathway displays and display options for PathBank 2.0 and these are described later in the Pathway Improvements section.

For users opting to browse the **Primary Pathways** or **TOPP**, a separate multi-page table with four columns is presented that lists the: (i) *PathBank ID*; (ii) *Primary Pathway Name*; (iii) *Primary Pathway Class* and (iv) a link to the *Primary Pathway View* with the number of associated secondary pathways listed below the 'View Pathway' button. The table is sortable by *PathBank ID*, *Pathway Name*, *Pathway Class* and Number of secondary pathways by clicking on the up/down arrows in the column titles. The TOPP Table is also filterable by species and pathway type in a manner similar to that described for the Pathway Table. Primary pathways with more than one secondary pathway are displayed as a primary pathway with a clickable name. These hyperlinked pathways lead to related pathways in pop-up boxes or links. For lipid pathways, only the generic version is listed; specific lipid pathways are accessible through the generic lipid's pop-up.

For users opting to browse **Compounds** or **Proteins**, similar multi-page tables display specific compounds and proteins alphabetically, with four columns: (i) *Compound* or *Protein ID* (with a thumbnail image); (ii) *Compound* or *Protein Description*; (iii) *Pathway Class* (with hyperlinks) and (iv) associated *Pathways* (hyperlinked and also displaying total number of pathways). Clicking a compound's 'View' button directs to the relevant 'MetaboCard,' defaulting to HMDB if no organism is chosen. Clicking a protein's 'View' button leads to the corresponding 'UniProt Cards.' The tables are sortable by *Compound/Protein ID, Compound/Protein Name, Pathway Class Name* and number of Pathways by clicking the up/down arrows at the top of the table. Compounds and proteins are filterable by species and *Pathway Class* through dropdowns, and a search box aids additional filtering by name.

In addition to these improved browsing functions, PathBank 2.0 also offers extensive search functions under the 'Search' tab including: (i) An *Advanced Text Search* (with instructions): (ii) the *ChemQuery Structure Search*; (iii) a *Molecular Weight Search* and (iv) a *Sequence Search*. The operations, instructions and applications for these search tools is identical to what was described in the original PathBank paper (5). New to PathBank 2.0 is the 'Analyze' tab. Three pathway analysis options are offered under this tab: (i) over-representation analysis; (ii) Path-MAP analysis and (iii) pathway enrichment analysis. These pathway analysis functions are described in more detail in a later section entitled New Pathway Analysis Tools. The 'About' tab in PathBank 2.0 provides additional information about PathBank, its associated release notes, the required citations, database statistics, the PathBank 2.0 taxonomy, the PathBank 2.0 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 2.0 pathways. The 'Download' tab for PathBank 2.0 allows users to download pathways, metabolite names, and protein names (in CSV or TSV file format) as well as all of PathBank 2.0′s pathways in BioPAX, SBGN, SBML and PWML format.

Database implementation

PathBank 2.0 was developed using the Ruby on Rails web framework (http://rubyonrails.org, version 4.2.0). This framework employs a MariaDB relational database (https://www.mariadb.com, version 5.5.56) to comprehensively manage various aspects of the pathway data. This includes details such as entity relationships, external references, comprehensive descriptions, visualization specifications, and even chemical structures. The design of PathBank 2.0 follows the Model-View-Controller (MVC) architecture, an approach that elegantly separates internal data logic from user input and data presentation. This separation ensures a modular and organized structure to the application's codebase. The core information stored within PathBank 2.0 is extracted in a dynamic manner and then seamlessly transformed into visually appealing web pages through PathBank's HTML interface responder. This allows users to interact with the data easily and in a user-friendly, informative manner. PathBank 2.0 is made accessible through a dedicated server hosted on Digital Ocean. This server has 8 vCPUs that ensure robust processing capabilities, 320 GB of disk space to accommodate extensive data storage, and 16 GB of RAM to ensure smooth performance. Additionally, an extra 500 GB of storage space is allocated on an Amazon S3 storage facility, further enhancing PathBank's ability to house and manage vast amounts of pathway-related information.

Improved database curation and quality assurance

All members of PathBank's curation team were required to have at least undergraduate degrees or senior-level undergraduate courses in bioinformatics, biochemistry, pharmacology and/or molecular biology. Training in pathway illustration was standardized and was conducted over a two-week period by the lead curator(s). This training included direct mentoring, standard operating protocol (SOP) instruction, practice pathway illustration, peer support and tutorials. All pathways generated by PathBank's curation team followed established SOPs, checklists, and PathBank 2.0 style guides for uniformity (5,6,8). The PathBank 2.0 style guide is available under the 'About' tab. To select pathways for a particular species or pathway category, the curation team conducted literature searches for unique metabolite/pathway classes, metabolic processes, and proteins. Existing pathway diagrams from public databases were reviewed, compared to internal references, and discussed. New pathways were identified, selected, hand-sketched, and then illustrated on-line using PathWhiz by members of the curation team. Each pathway is independently evaluated by a second curation team member for adherence to style guides. Pathway reviews by senior members of the team were conducted on a biweekly basis. Only compliant pathways were propagated or replicated into PathBank 2.0 using PathWhiz. This quality control applies to primary pathways which serve as templates for replication. Spot checks were used to ensure consistency, and system-wide checks were used to verify image appropriateness. Primary pathways were more frequently reviewed to ensure they aligned fully with PathBank's style guide. Typically, a single primary pathway would take a skilled curator between 8–20 hours to research, render and complete while a single secondary pathway would take 15 minutes to 1 hour to research, render and complete. It is important to note that PathBank's pathways undergo constant minor layout improvements and revisions. Minor corrections occur without formal announcements. Significant changes (>10% in components) are noted in descriptions with modified update dates.

Pathway additions

Since its last release in 2020, the PathBank team has put in thousands of hours to add more than 605 000 pathways for 10 model organisms to the PathBank database. Some of the most important new pathways are *Primary Pathways*, meaning that they are biologically or biochemically distinct pathways that describe an important or general molecular/physiological process. *Primary Pathways* require extensive literature research, careful pathway design, many hours of rendering and even more hours of internal/external vetting of the curation team. In total, 6951 *Primary Pathways* were added to PathBank 2.0, which represents a 400% increase in the number of *Primary Pathways* originally present in PathBank 1.0. Some of these newly added *Primary Pathways* as well as previously existing *Primary Pathways* were then replicated or propagated (with suitable modifications and updates based on new chemical knowledge) to produce another 598 408 *Secondary Pathways*.

For the past two years, a major focus for PathBank's curation team has been on expanding the content of four major pathway classes that were under-represented in PathBank 1.0: (i) drug action or drug mechanism pathways; (ii) drug metabolism pathways; (iii) uremic toxin pathways and (iv) physiological pathways. This is because there has been a significant dearth of these kinds of pathways in most pathway databases. This under-representation has, unfortunately, been leading to inconsistent and often incorrect interpretations of the results of many metabolomics assays (35,36). As might be expected, most of these new pathway additions were specific to humans, although a number of pathways were propagated to other mammalian species (*B. taurus, R. norvegicus, M. musculus, etc.*). In total 1440 primary drug action pathways, 5254 primary drug metabolism pathways, 65 primary uremic toxin pathways, 50 primary disease mechanism and 142 primary physiological pathways were generated and added to PathBank 2.0. A portion of these *Primary Pathways* (mostly lipid pathways) were further replicated with appropriate modifications to produce another 142 588 *Secondary Pathways*, which were specific to *Homo sapiens* alone. In addition, some of these were propagated (transferred) to other organisms to bring the total number of added *Secondary Pathways* in PathBank 2.0 to 598 408.

While the drug action pathways required the most manual effort, the drug metabolism pathways are of particular interest as these were generated through a new computational approach for PathBank and PathWhiz that may become more widely used in future versions of this database. In particular, experimentally determined drug metabolism reaction data was collected from DrugBank (26) and further supplemented with drug metabolism data generated (via machine learning) from BioTransformer (37,38) for drugs that appeared to be missing key drug metabolites (26). Because information about the enzymes, organs and body compartments could be extracted from these data, it was possible to automatically generate thousands of richly annotated and remarkably complete drug metabolism pathways. Pathways that use

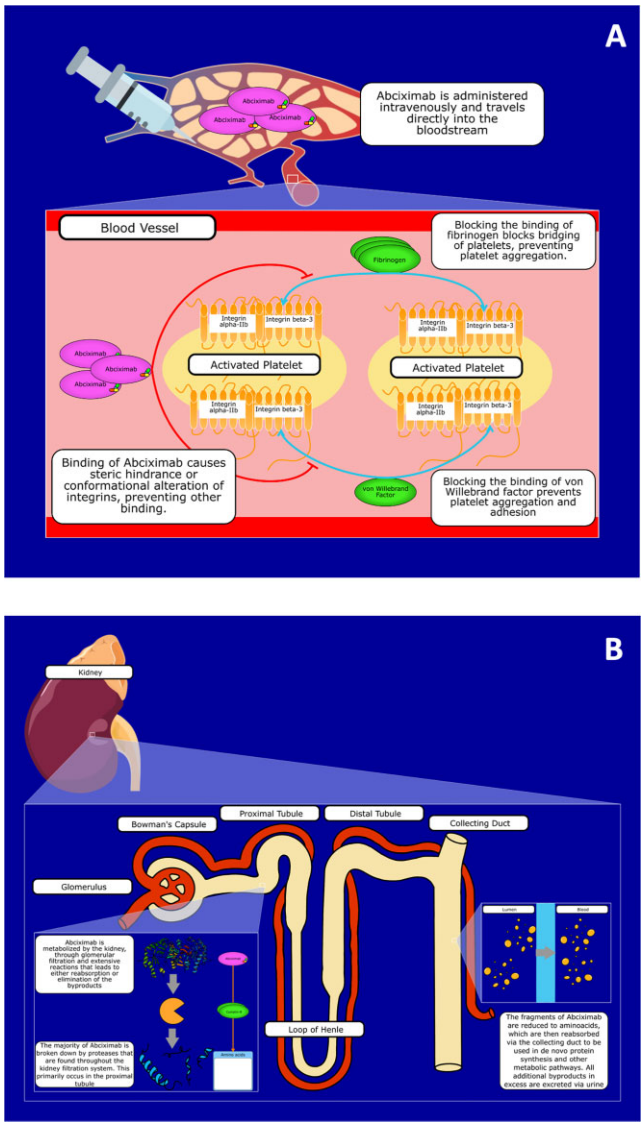


Figure 1. (**A**) The drug action (mechanism of action or MOA) and (**B**) drug metabolism (absorption, distribution, metabolism and excretion or ADME) pathway for the drug Abciximab.

BioTransformer predictions have been labeled as predicted, so that users are aware that the data may not be as reliable as the experimentally confirmed pathways. Examples of both the drug action and drug metabolism pathway for the drug Abciximab are shown in Figure 1.

In addition to these *Primary Pathways*, the PathBank team also took advantage of new lipid composition data appearing in the latest release of the HMDB 5.0 (5) to help replicate another 138 952 lipid biosynthesis pathways, primarily among triglycerides, phospholipids, ceramides and acylcarnitines. Overall, across the different model species in PathBank, pathway additions to PathBank 2.0 were most concentrated on *H. sapiens* (140 004 new pathways), followed by 126 749 new pathways for *B. taurus*, 126 530 new pathways for *R. norvegicus*, 126 530 new pathways for *M. musculus*, 7 new pathways for *D. melanogaster*, 11 new pathways for *A. thaliana*, 437 new pathways for *S. cerevisiae*, 22 new pathways for *E. coli*, and 83 new pathways for *P. aeruginosa*.

Pathway improvements

The addition of so many diverse *Primary Pathways* to such a wide collection of species within PathBank led to two major initiatives to improve pathway quality in PathBank 2.0. In particular, the PathBank team focused on: (i) pathway remediation and (ii) pathway enhancement. Pathway remediation involved assessing the quality of previously illustrated PathBank pathways and improving the layout, logic, content, iconography and quality of old pathways. Pathway enhancement involved improving the quality, layout, logic, content and iconography of newly added pathways. Both processes (remediation and enhancement) informed each other and both processes required additional curator training, regular team meetings as well as constant updating of SOPs, the PathBank style guide, image or icon libraries and the PathWhiz pathway illustration software.

One example of a key pathway improvement or enhancement involved the inclusion of standard organelle images, standard chromosome depictions and standard cell structures for appropriate cell types (i.e. different cell walls for different types of prokaryotes, the nucleus and mitochondria for eukaryotes). This was introduced for all new pathways and back-propagated to all previously illustrated pathways. Another example included the establishment of minimum sizes for organs, tissues, organelles and metabolites and proteins in all pathway illustrations. Again, this was applied to previous pathways as part of the remediation effort and continued with all newly added pathways. A third example involved the creation of many new icons or images to depict more diverse organelles (the endoplasmic reticulum, vacuoles, Golgi apparatus, vesicles, chloroplasts, peroxisomes, muscle filaments, etc.), more cell types (red blood cells, B cells, T cells, macrophages, neurons, fibroblasts, astrocytes, glia, muscle cells, bacteria and viruses) and more organs or tissues (brain, kidney, liver, kidney substructures, muscle, the gastrointestinal tract, pancreas, thymus, bone, adipose tissue, skin or epithelium and blood vessels). An example of how the addition of these icons and images improved the content and quality of PathBank pathways is shown in Figure 2, which illustrates the process by which the COVID-19 virus causes viral sepsis.

Other pathway improvements or remediation efforts undertaken by the PathBank curation team involved adding more transport proteins to pathways (where possible) to indicate how molecules entered or left cells/tissues, adding more 'Zoom' boxes to pathways to illustrate different scales of activity and adding more text annotations to pathways to clarify complex processes or events that might occur in tissues, organs or organelles. For instance, reactions or processes that occur in the mitochondria or the nucleus are shown through a 'Zoom' box or expansion box that is illustrated within the cell itself. A PathBank 'Zoom' box typically has a triangular perspective diagram with the vanishing point emanating from the center of the organelle which leads to the full (rectangular) view of the inside of the organelle. These expanded or zoomed-out pathways maintain the same color background as used to depict the organelle. Therefore, a pink background indicates a pathway localized to the mitochondrion while a yellow background indicates a pathway localized to the nucleus. A number of pathways were also reorganized to better illustrate the sequence of events or the correct positioning of organs involved in the biological/physiological process. Likewise, because PathBank covers a wide number of species, new

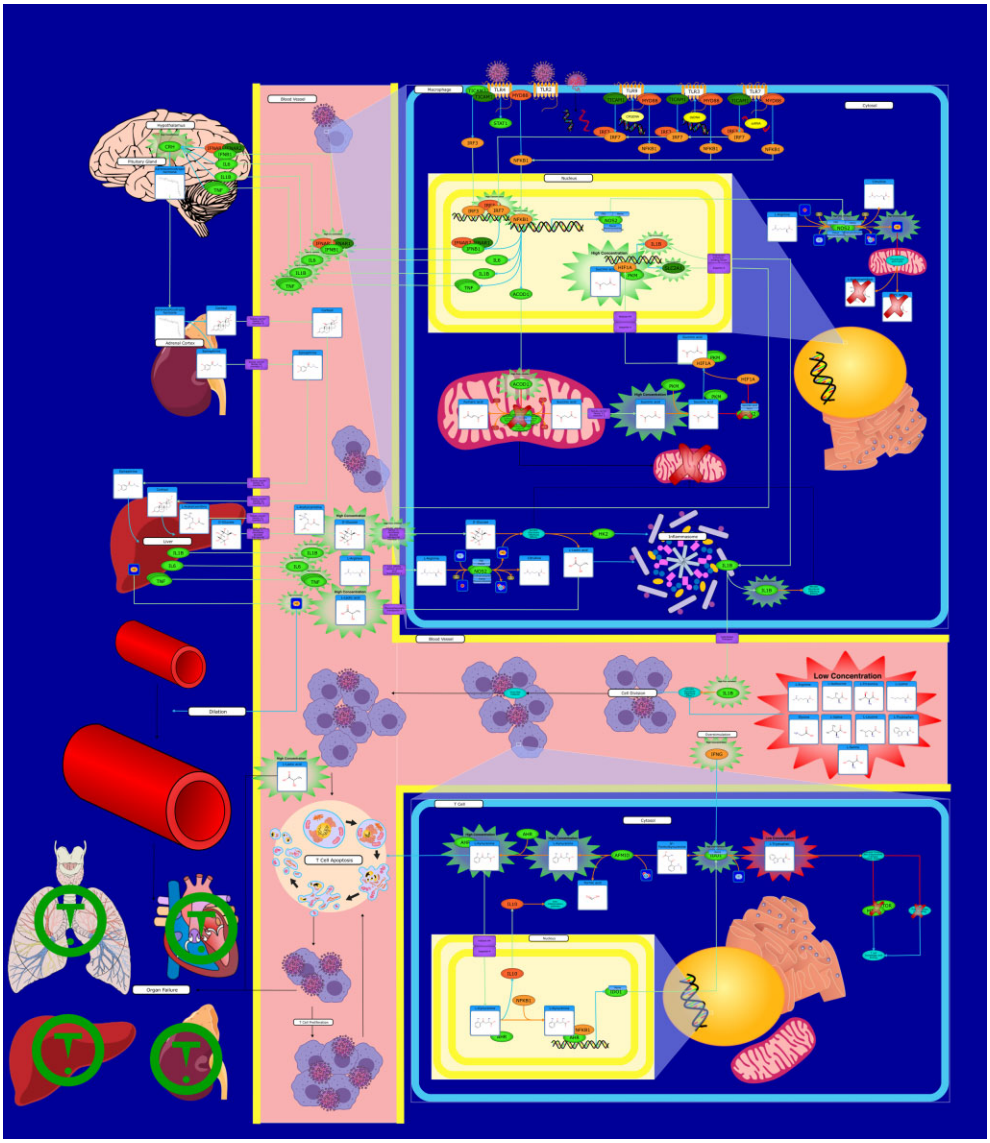


Figure 2. Pathway illustrating how the COVID-19 virus can cause viral sepsis. This showcases the additional icons and images that have been added to improve the content and quality of PathBank pathways.

images and new icons were added or edited to more appropriately reflect the correct shape of organs (brain, liver, gut, etc.), body parts and other physiological aspects of the organism being depicted.

A major focus for the past three years was on improving the compatibility of PathBank's pathway images for use in slides, posters and publications. Specifically, all pathway images and pathway icons (cell organelles, organs, tissues, body parts, etc.) now exist as both full color and grayscale images. Using the 'Settings/Display" associated with each PathBank 2.0 image, users can modify and customize the pathway image to suit their publication needs. Users may select a dark blue (default) background or a white background or between the full colour (full detail), grey-scale (full detail) and simplified KEGG-like (black and white) pathway representation. Likewise, for publication quality figures, users may use the 'Downloads' option associated with the pathway and select one of two options: 'Large Font SVG Image' for detailed pathway figures, or 'Simple Large Font SVG Image' for a less cluttered image. Text scaling (font size) has been improved so that the names of pathway components and labels are now far more visible. Furthermore, users can choose to display the names as abbreviations to further increase text size. Likewise, the placement of pathway nodes corresponding to metabolites/compounds in the KEGG-like pathway diagrams has been improved to reduce text overlap. The KEGG-like pathway diagrams also include cell, organelle and organ information to help maintain biological context. These improvements should make all of PathBank 2.0's pathway diagrams much more useful for publication purposes.

New pathway analysis tools

As noted earlier, a new addition to PathBank 2.0 is the pathway 'Analysis' option. Clicking on the 'Analysis' tab produces a pull-down menu with three Analysis options: (i) Over-Representation Analysis (ORA); (ii) Path-MAP Analysis and (iii) Pathway Enrichment Analysis (PEA). Both ORA and PEA are new to PathBank 2.0, while Path-MAP has been updated from PathBank 1.0. For this release of PathBank 2.0, ORA

is limited to handling metabolite data only. Fundamentally, ORA works by identifying metabolic pathways or metabolite sets that have a higher overlap with a set of measured metabolites of interest than expected by chance. To perform ORA with PathBank 2.0, three inputs are required: the choice of organism, a list of experimentally measured, significant metabolites of interest, and a reference metabolite/pathway set. After choosing the organism (via a pull-down menu), users must provide a list of significantly altered compounds in the form of compound names or compound identifiers. Information about the allowable identifiers or names, as well as appropriate examples, is given in the ORA page under 'Allowed Nomenclature'. Example data sets for running ORA are provided under the 'Example' button. Once the data set is uploaded, users must select a metabolite set library from the set of offered pathway libraries provided through a pull-down menu. PathBank 2.0 offers several libraries including the full PathBank metabolite library, a Disease metabolite library, a Signaling metabolite library, a Physiological metabolite library and several biofluid-specific metabolite libraries. Once the pathway library has been selected, users can press the 'Calculate ORA' button to generate the ORA network diagram, bar graphs and tables. The ORA network diagram displays the pathway names and their connectivity with the size of the nodes indicating their significance. All nodes in the network diagram are clickable and reveal additional details about the named pathways. The bar graph displays the level of pathway enrichment as determined by the ORA algorithm and the table provides statistical measures of the significance. The ORA algorithm uses a hypergeometric test based on the cumulative binomial distribution to determine if input metabolites are found among pathway metabolites more often than would be expected by chance. Additional details about ORA calculations and their interpretation are available through various tutorials written for MetaboAnalyst (30).

The Path-MAP Analysis, which was previously described in the original PathBank paper (5), supports metabolomic, proteomic and transcriptomic analysis. It allows users to enter lists of compound names, protein names, compound identifiers, protein identifiers, or gene identifiers that have been identified as being significant via experimental studies in metabolomics, proteomics or transcriptomics. Information about the allowable identifiers or names, as well as appropriate examples, is given in the Path-MAP page under 'Allowed Nomenclature'. Example data sets for running Path-MAP are provided under the 'Example 1' button. Users must provide the list of biomolecules and choose the organism (from a pull-down menu) and press 'Run Path-MAP'. Path-MAP then maps these biomolecules to the pathways for the specified organism. 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. Through Path-MAP it is also possible to enter biomolecular lists with concentration or relative concentration data, as might be obtained from a typical proteomics, transcriptomics, or metabolomics experiment. Example data sets for running Path-MAP using this option are provided under the 'Example 2' button. When this option is run, it produces an organism-specific pathway annotated and coloured with the corresponding metabolite/protein/transcript concentrations according to a yellow (low)-red (high) concentration gradient. More details about Path-MAP's functions are described in the previous release of PathBank (5).

The Pathway Enrichment Analysis function is new to PathBank 2.0 and offers users the ability to enter lists of significantly altered compounds, genes or proteins (from metabolomics, transcriptomics or proteomics experiments) from a standard comparative study (case versus control) and to generate colorful, informative pathway enrichment plots. Users must provide the list of experimentally measured metabolites, genes or proteins, their normalized/scaled values and their associated *P*-values. These types of 'omics' datasets can be easily generated from standard 'omics' analysis tools or software packages. As with PathBank's other analysis tools, users may provide compound names, protein names, compound identifiers, protein identifiers or gene identifiers. Information about the allowable identifiers or names, as well as appropriate examples, is given in the Pathway Enrichment page under 'Allowed Nomenclature'. Once the appropriate dataset is selected and submitted, users must select the organism and the pathway library and then press the 'Calculate Pathway Impact" button to generate the Pathway Impact diagram. Example data sets are provided under the 'Examples' button. Users may select from a number of organism-specific pathway libraries covering many PathBank pathway categories, subcategories and classes. Screenshots illustrating how to use pathway enrichment analysis in PathBank 2.0 are shown in Figure 3. All nodes in the pathway impact graph are clickable (Figure 3B) and reveal additional details about the pathways through tables that provide statistical measures of significance (Figure 3C). Pathway impact is calculated using a classic measure of network topology called out-degree centrality. Briefly, if each pathway is a network, the nodes are metabolites, and out-degree centrality is the total number of connections (degrees) that originate from a given node. The sum of out-degree centrality for matching metabolites in the pathway is then normalized by the sum for all metabolites in the pathway to provide the impact factor. Additional details about Pathway Impact calculations and interpretations are available through MetaboAnalyst (30).

Conclusion

We believe that PathBank 2.0 brings a number of important improvements and enhancements that should make the database much more useful and far more appealing to a larger cross-section of users. Specifically, PathBank 2.0 now contains thousands of new primary pathways and hundreds of thousands of new secondary pathways spanning a range of pathway types (such as drug mechanism and drug action pathways) that had not previously been covered by any pathway database. PathBank 2.0 also brings significant improvements to the quality and content of its pathway diagrams and pathway content. These include more informative diagrams, more complete depictions and a wider array of subcellular and supracellular images. In addition to the improved images, each protein and metabolite depicted in PathBank 2.0 has associated physiological information including the tissue, cell type and subcellular location. This information can be found in the pop-ups that appear when clicking on a protein or a metabolite in a PathBank 2.0 pathway display. A significant effort was undertaken to make all of PathBank 2.0′s pathway images far more slide-friendly and manuscript-compatible by improving the layout, increasing the text size and enhancing the quality of the grey-scale and black-and-white pathway diagrams. We have also improved PathBank's pathway

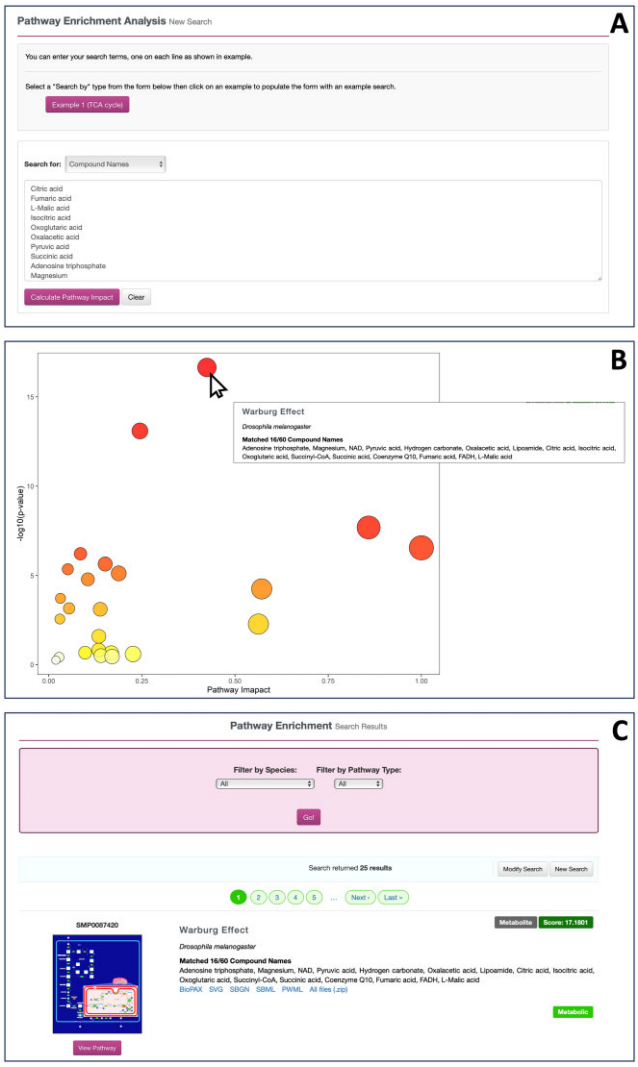


Figure 3. Screenshots illustrating how to use Pathway Enrichment Analysis (PEA). (**A**) View of the PEA search screen with example data entered. (**B**) A Pathway Impact plot of the results from the PEA. The Pathway Impact plot is interactive and when individual dots are hovered over, pathway information is provided. (**C**) Shown is the results table which gives more pathway information.

taxonomy and added new tools to support better pathway filtering, selection and display. Finally, we have added or enhanced a number of pathway analysis tools for visualizing and calculating pathway enrichment and pathway impact. Many other minor improvements and updates to the content, the interface, and general performance of the PathBank website have also been made. Overall, we believe these upgrades and updates should greatly enhance PathBank's ease of use and its potential applications for interpreting a wide range of omics data.

Data availability

PathBank 2.0 is FAIR compliant. An extensive and well-annotated data download section is also provided with most data pathways, metabolite names, and protein names available in standard CSV and TSV file formats and all pathways in BioPAX, SBGN, SBML and PWML formats. The data in PathBank 2.0 are released under the Creative Commons (CC) 4.0 License Suite according to the Attribution BY and Non-Commercial NC licensing conditions.

Acknowledgements

We wish to thank Dr Marcia LeVatte for her help in preparing and proofreading this manuscript.

Funding

Canadian Institutes of Health Research (CIHR); Natural Sciences and Engineering Research Council (NSERC) Alliance Program; Alberta Innovates (via the Campus Alberta Small Business Engagement Program - CASBE); OMx Personal Health Analytics Inc.; Genome Alberta, a division of Genome Canada. Funding for open access charge: CIHR; NSERC Alliance; Alberta Innovates CASBE; Genome Canada; cash contributions from OMx Personal Health Analytics Inc.

Conflict of interest statement

None declared.

References

1. Pinu,F.R., Beale,D.J., Paten,A.M., Kouremenos,K., Swarup,S., Schirra,H.J. and Wishart,D. (2019) Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites*, **9**, 76.
2. Heckendorf,C., Blum,B.C., Lin,W., Lawton,M.L. and Emili,A. (2023) Integration of metabolomic and proteomic data to uncover actionable metabolic pathways. *Methods Mol. Biol.*, **2660**, 137–148.
3. Wanichthanarak,K., Fahrmann,J.F. and Grapov,D. (2015) Genomic, proteomic, and metabolomic data integration strategies. *Biomark Insights*, **10**, 1–6.
4. Cavill,R., Jennen,D., Kleinjans,J. and Briedé,J.J. (2016) Transcriptomic and metabolomic data integration. *Brief. Bioinform*, **17**, 891–901.
5. Wishart,D.S., Li,C., Marcu,A., Badran,H., Pon,A., Budinski,Z., Patron,J., Lipton,D., Cao,X., Oler,E., *et al.* (2020) PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res.*, **48**, D470–D478.
6. Frolkis,A., Knox,C., Lim,E., Jewison,T., Law,V., Hau,D.D., Liu,P., Gautam,B., Ly,S., Guo,A.C., *et al.* (2009) SMPDB: the small molecule pathway database. *Nucleic Acids Res.*, **38**, D408–D487.
7. 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.
8. 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.
9. 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. *Journal of Visualized Experiments*, **2017**, 54869.
10. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
11. 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.
12. Caspi,R., Billington,R., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Midford,P.E., Ong,Q., Ong,W.K., *et al.* (2018) The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **46**, D633–D639.

13. Karp,P.D., Billington,R., Caspi,R., Fulcher,C.A., Latendresse,M., Kothari,A., Keseler,I.M., Krummenacker,M., Midford,P.E., Ong,Q., *et al.* (2019) The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform*, **20**, 1085–1093.
14. 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.
15. 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.
16. Pico,A.R., Kelder,T., Van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
17. 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.
18. 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.
19. 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.
20. Novère,N.L., Hucka,M., Mi,H., Moodie,S., Schreiber,F., Sorokin,A., Demir,E., Wegner,K., Aladjem,M.I., Wimalaratne,S.M., *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.
21. 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.
22. 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.
23. Breuer,K., Foroushani,A.K., Laird,M.R., Chen,C., Sribnaia,A., Lo,R., Winsor,G.L., Hancock,R.E.W., Brinkman,F.S.L. and Lynn,D.J. (2013) InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation. *Nucleic Acids Res.*, **41**, D1228–D1233.
24. Nishimura,D. (2001) A view from the web. BioCarta. *Biotech. Softw. Internet Rep.*, **2**, 117–120.
25. Wishart,D.S., Guo,A., Oler,E., Wang,F., Anjum,A., Peters,H., Dizon,R., Sayeeda,Z., Tian,S., Lee,B.L., *et al.* (2022) HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.*, **50**, D622–D631.
26. 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.
27. Wishart,D.S., Bartok,B., Oler,E., Liang,K.Y.H., Budinski,Z., Berjanskii,M., Guo,A., Cao,X. and Wilson,M. (2021) MarkerDB: an online database of molecular biomarkers. *Nucleic Acids Res.*, **49**, D1259–D1267.
28. Sherman,B.T., Hao,M., Qiu,J., Jiao,X., Baseler,M.W., Lane,H.C., Imamichi,T. and Chang,W. (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.*, **50**, W216–W221.
29. Xia,J. and Wishart,D. (2016) Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Curr. Protoc. Bioinformatics*, **55**, 14.10.1–14.10.91.
30. Chong,J., Wishart,D.S. and Xia,J. (2019) Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Curr. Protoc. Bioinformatics*, **68**, e86.
31. Pang,Z., Chong,J., Zhou,G., De Lima Morais,D.A., Chang,L., Barrette,M., Gauthier,C., Jacques,P.É., Li,S. and Xia,J. (2021) MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res.*, **49**, W388–W396.
32. Xia,J. and Wishart,D.S. (2011) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.*, **6**, 743–760.
33. Chong,J., Soufan,O., Li,C., Caraus,I., Li,S., Bourque,G., Wishart,D. and Xia,J. (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.*, **46**, W486–W494.
34. Petri,V., Jayaraman,P., Tutaj,M., Hayman,G.T., Smith,J.R., De Pons,J., Laulederkind,S.J.F., Lowry,T.F., Nigam,R., Wang,S.J., *et al.* (2014) The pathway ontology - updates and applications. *J. Biomed. Semantics*, **5**, 7.
35. Wieder,C., Frainay,C., Poupin,N., Rodríguez-Mier,P., Vinson,F., Cooke,J., Lai,R.P.J., Bundy,J.G., Jourdan,F. and Ebbels,T. (2021) Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. *PLoS Comput. Biol.*, **17**, e1009105.
36. Chen,Y., Li,E.M. and Xu,L.Y. (2022) Guide to metabolomics analysis: a bioinformatics workflow. *Metabolites*, **12**, 357.
37. Wishart,D.S., Tian,S., Allen,D., Oler,E., Peters,H., Lui,V.W., Gautam,V., Djoumbou-Feunang,Y., Greiner,R. and Metz,T.O. (2022) BioTransformer 3.0—a web server for accurately predicting metabolic transformation products. *Nucleic Acids Res.*, **50**, W115–W123.
38. Djoumbou-Feunang,Y., Fiamoncini,J., Gil-de-la-Fuente,A., Greiner,R., Manach,C. and Wishart,D.S. (2019) BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminform.*, **11**, 2.

Received: September 15, 2023. **Revised:** October 19, 2023. **Editorial Decision:** October 20, 2023. **Accepted:** October 31, 2023