Computational and Structural Biotechnology Journal 20 (2022) 2280–2291

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Data article

OncoboxPD: human 51 672 molecular pathways database with tools for activity calculating and visualization

Marianna A. Zolotovskaia [a,b,c,*], Victor S. Tkachev [a], Anastasia A. Guryanova [a,d], Alexander M. Simonov [e], Mikhail M. Raevskiy [e], Victor V. Efimov [e], Ye Wang [f], Marina I. Sekacheva [e], Andrew V. Garazha [a], Nicolas M. Borisov [b], Denis V. Kuzmin [b], Maxim I. Sorokin [a,b,d,h], Anton A. Buzdin [b,e,g,h]

[a] Omicsway Corp., Walnut, CA 91789, USA
[b] Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, 141701, Russia
[c] Pirogov Russian National Research Medical University, Moscow 117997, Russia
[d] I.M. Sechenov First Moscow State Medical University, Moscow 119991, Russia
[e] World-Class Research Center "Digital biodesign and personalized healthcare", Sechenov First Moscow State Medical University, Moscow, Russia
[f] Clinical Laboratory, Qingdao Central Hospital, The Second Affiliated Hospital of Medical College of Qingdao University, Qingdao, China
[g] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow 117997, Russia
[h] PathoBiology Group, European Organization for Research and Treatment of Cancer (EORTC), Brussels, Belgium

ARTICLE INFO

Article history:
Received 1 March 2022
Received in revised form 5 May 2022
Accepted 5 May 2022
Available online 10 May 2022

Keywords:
Molecular pathway database
Interactomics
Protein–protein interactions
Metabolomics
Pathway activation level
Pathway visualization

ABSTRACT

OncoboxPD (Oncobox pathway databank) available at https://open.oncobox.com is the collection of 51 672 uniformly processed human molecular pathways. Superposition of all pathways formed interactome graph of protein–protein interactions and metabolic reactions containing 361 654 interactions and 64 095 molecular participants. Pathways are uniformly classified by biological processes, and each pathway node is algorithmically functionally annotated by specific activator/repressor role. This enables online calculation of statistically supported pathway activation levels (PALs) with the built-in bioinformatic tool using custom RNA/protein expression profiles. Each pathway can be visualized as static or dynamic graph, where vertices are molecules participating in a pathway and edges are interactions or reactions between them. Differentially expressed nodes in a pathway can be visualized in two-color mode with user-defined color scale. For every comparison, OncoboxPD also generates a graph summarizing top up- and downregulated pathways.

1. Introduction

Molecular pathway research is a rapidly growing field that is developing exponentially since the emergence of high-throughput microarray, next generation sequencing, proteomic technologies, and supporting bioinformatic tools [1–7]. Molecular pathways describe certain biological processes at the molecular level and include up to several hundred different molecular participants [7]. Molecular pathways most frequently are understood as networks of protein–protein interactions, or biochemical reactions, or their combinations [6].

At present, multitude of molecular interactions was documented, and thousands of molecular pathways were reconstructed and published [5,6]. The molecular interactions and reactions can be detected both experimentally and by using artificial intelligence methods [8,9].

Several high-throughput databases of molecular interactions and molecular pathways became available over the past decades, including Pathway interaction database [10], QIAGEN SABiosciences [11], Pathway Studio [12], Reactome [13], Kyoto Encyclopedia of Genes and Genomes (KEGG) [14], Signaling Pathways Integrated Knowledge Engine (SPIKE) [15], Metacyc [16], HumanCyc [17], and PathBank [6]. Some of these databases select pathways by their function, such as HumanCyc collection of human metabolic pathways[17], or SynSysNet collection of synaptic

Abbreviations: ARR, activation/repressor role coefficient; BEL, Biological Expression Language; KEGG, Kyoto Encyclopedia of Genes and Genomes; PAL, pathway activation level; SPIKE, Signaling Pathways Integrated Knowledge Engine; SPOKE, Scalable Precision Medicine Open Knowledge Engine; VM, virtual machine.

* Corresponding author.
E-mail address: zolotovskaya@oncobox.com (M.A. Zolotovskaia).

https://doi.org/10.1016/j.csbj.2022.05.006

protein–protein interactions[18]. Some databases propose specific classification terms, e.g. *metabolism*, *genetic information processing*, *environmental information processing*, *cellular processes*, *organizational systems*, *diseases*, *drug development* categories of pathways in KEGG [14], or *metabolic*, *drug action*, *drug metabolism*, *disease*, *signaling*, *physiological* categories in Pathbank [6]. Alternatively, four molecular functions-based categories of *signaling*, *metabolic*, *cytoskeleton*, and *DNA repair* were recently proposed for the pathway classification [19].

However, to our knowledge all these pathway functional classification types are not algorithmically standardized, thus leading to non-uniformly functionally annotated groups of pathways. Importantly, pathways frequently describe complex molecular processes, and one pathway may be related to several functional categories. Furthermore, there is a lack of uniform annotation for the individual molecular participants of intracellular pathways (e.g. pathway nodes) in terms of their overall functional implication in the activation of a pathway.

With the increasing amount of OMICS data, new instruments are needed for the unbiased accurate high-throughput analysis of molecular pathways, including their annotation and visualization. At present there are several molecular pathway visualization tools [20] which can be grouped in two types. In the first type, the outputs are static pathway images not depending on the user's custom data. Such images are useful for illustrating functional interactions under study without integrating them with the molecular expression data [21]. In the second type tools, interactive figures can be obtained [22]. For example, KEGG software can use gene expression data to accentuate specific individual nodes on a pathway [14], and Reactome software can highlight relevant pathways on an overall interaction map [13]. PathBank software can process the input concentrations of metabolites while returning their adjusted visualization on an interactome graph [6].

However, next-generation molecular pathway analysis requires not only visualization, but also calculation of numeric pathway characteristics (e.g. extents of their up/downregulation) based on the user experimental data. Currently, there is a lack of uniformly assigned weights and coefficients reflecting each individual gene product/node role in the activation of a pathway under study.

Recently, an algorithmic approach was proposed for assigning of activation/repressor roles to the pathway components that was applied for the automatic annotation of 3040 pathways [23]. These roles were translated into node-specific numeric indexes necessary for calculating pathway activation levels (PALs).

Here, we present Oncobox pathway databank that accumulates 51 672 uniformly processed human molecular pathways extracted from different source databases. Superposition of all pathways formed interactome graph of protein–protein interactions and metabolic reactions containing 361 654 interactions and 64 095 molecular participants. All pathways were functionally classified by their main underlying biological processes according to Gene Ontology (GO) tree. Each pathway node was algorithmically functionally annotated by specific activation/repressor role index. This enables direct calculation of pathway activation levels (PALs) using human RNA/protein expression profiles. With the Web-based bioinformatic tool or downloadable *oncoboxlib* Python library, user can analize custom expression data to assess PALs in the samples of interest compared to the built-in or custom set of controls, and statistically evaluate differentially regulated pathways. Each pathway can be visualized both as static or dynamic graph, where vertices are molecules participating in a pathway and edges are interactions or reactions between them. Differentially expressed nodes in a pathway can be visualized in two-color mode with user-defined color scale. Online version of Oncobox PD is freely available at https://open.oncobox.com.

2. Materials and methods

OncoboxPD utilizes virtual machines (VMs) located in Microsoft Azure cloud. All VMs run on Ubuntu 20.04.3 LTS. On the entry node, HTTP web server nginx is installed that is responsible for processing user requests (2 vCPUs, 8 GB RAM, D2s size according to Microsoft classification).

HTTP server and the processes running in Python communicate through Web Server Gateway Interface (uWSGI 2.0.19 implementation).

JavaScript web application (running in the browser of a user) interacts via REST API with the backend. On the server we have a Python 3.7 application build on top of Django framework. A single instance of PostgreSQL 13.3 is used as a data storage.

Calculation tasks are sent via RabbitMQ queue for processing to compute optimized nodes (4–8 vCPUs, 8–16 GB RAM, F4s-F8s size according to Microsoft classification). These settings can be dynamically scaled up or down depending on current workload. Once the calculation is complete, the results are returned to the main node and made available to the user.

3. Results

3.1. Database content

Oncobox pathway databank (OncoboxPD) includes 51 672 human molecular pathways extracted from seven pathway knowledge bases: Biocarta [24], KEGG [25], HumanCyc [17], Qiagen [11], NCI [10], Reactome [26] and PathBank [6]. The data were collected by combining manual and automatic parsing and curation of these source datasets. The processed pathways are stored in OncoboxPD with their original names in uniform format. Where possible, information on pathway participants and their interactions was extracted and catalogued (Table 1).

Since most of the initial pathway formats had different nomenclatures for designation of genes, proteins, metabolites, and others relevant items, in OncoboxPD we introduced uniform nomenclature according to the following rules. All genes and their products are named according to HGNC nomenclature [27], version from 17 July 2017; metabolites are annotated by full names and also linked to IDs from Chemical Abstract Service (CAS) [28], Chemical Entities of Biological Interest (ChEBI) [29], The Human Metabolome Database (HMDB) [30], PathWiz [31] and DrugBank [32] resources, where available.

In addition to full-size pathways, we also included so-called micropathways, where such micropathway is a sub-graph of an existing pathway that contains major effector node and first to third order neighbor nodes. The micropathways were generated automatically using our previously published algorithm [23].

In the case of algorithmic parsing, the pathway connectivity data were controlled manually. The team implementing expert curation could edit algorithmically assigned node or component names, interaction marks and node composition to avoid artifacts, duplicates, and false interpretations. Pathway size (number of participants in a pathway) varied substantially within and among original datasets. In Biocarta, pathway size varied from 2 to 58 (average 16), in KEGG - from 1 to 284 (43), in HumanCyc - from 4 to 155 (26), in PathBank - from 2 to 217 (31), in NCI - from 2 to 180 (20), in Qiagen - from 2 to 709 (58), in Reactome - from 3 to 831 (36), Supplementary Figure S1.

Five source pathway datasets (Biocarta, KEGG, Qiagen, NCI, and Reactome) had only gene products as the pathway nodes, and two datasets (PathBank and HumanCyc) also contained metabolites as the interactors.

Table 1
Composition of source pathway knowledge bases.

Source pathway knowledge base	Number of pathways*	Number of gene products*	Number of metabolites*	Type of interactions comprized
Biocarta 1.2	337	1 082	–	protein–protein
KEGG 1.2	288	4 345	–	protein–protein
HumanCyc 1.0	300	980	1 040	protein–protein, biochemical reactions, transport
NCI 1.2	775	2 214	–	protein–protein
Qiagen 1.4 (SABiosciences)	379	2 493	–	protein–protein
Reactome 1.3	945	6 105	–	protein–protein
Pathbank 1.0	48 648	1 405	55 571	protein–protein, biochemical reactions, transport
Total	51 672	9 117	56 596	protein–protein, biochemical reactions, transport

*Numbers are shown for unique items only.

The percentage share of metabolic components per pathway was in the range of 0–94 (mean 74%) for the pathways from Pathbank, and 6–95 (mean 70%) for the pathways from HumanCyc database.

All processed pathway datasets are available for download as both "*.xlsx*" files for each separate pathway, and as combined "*.csv*" files for the whole dataset.

3.2. Pathway architecture

In OncoboxPD, each molecular pathway is implemented as a graph of interactions between its molecular participants. The graph edges are interactions between the nodes, such as protein–protein interactions, biochemical reactions, transport to different cellular components, assembly or disruption of molecular complexes, and other processes. Standard edge types in OncoboxPD are "activation", "inhibition" and "other" (Table 2). Each edge on the graph is directed, and when the underlying interaction is reversible, then two oppositely directed edges are placed between the interactors.

Graph vertexes (nodes) represent formal functional units of a pathway, or several molecular pathways that are interconnected on the graph (Table 3). Every node may include one or few components. For example, when a node stands for molecular complex, then it comprises few components which can be proteins (gene products) or non-protein molecules. In OncoboxPD, there is only one depth level for "molecular complex" nodes. Each node has an *activation/repressor role coefficient* (ARR) which characterizes its relation to the overall pathway physiological or molecular effect. ARR depends on role of an individual node is a pathway. In the current implementation, ARR = 1 means that node is activator, ARR = -1 is for repressor, ARR = 0 is for nodes with ambiguous or unclear function, and ARR = 0.5 or ARR = -0.5 is for nodes with rather activator or rather repressor functions, respectively.

Table 2
Type of pathway graph edges in OncoboxPD.

Initial edge type in Biopax or different format	Edge type in OncoboxPD format
Direct interaction, activation or inhibition	activation or inhibition, respectively
SubPathwayInteraction	other
ComplexAssembly	other
Molecular interaction(between participants from SubPathwayControl item) - activation or inhibition	activation or inhibition, respectively
Catalysis, activation or inhibition	activation or inhibition, respectively
Modulation (activation-allosteric, activation-nonallosteric, activation, inhibition-competitive, inhibition-other, inhibition-noncompetitive, inhibition-allosteric, inhibition-irreversible, inhibition)	activation or inhibition, respectively
Transport	activation, because it promotes further molecular interaction
BiochemicalReaction	activation, because it promotes further molecular interaction
Indirect	other
Compound	other
Others	other

ARR value of a node is translated into specific gene ARR, according to molecular functions that the respective gene product plays in all pathway nodes. This is mentioned because the same gene product can be involved in different nodes, sometimes with different ARRs. Thus, overall role of a gene product in a pathway must be weighted and characterized by an overall ARR value. This task can't be fulfilled effectively and unbiasedly by manual curation because of big number of pathways and their apparently high complexity. Thus, to uniformly annotate all pathways in OncoboxPD we used a recursive algorithm that was recently developed in our team to automatically annotate ARRs based on pathway graph architecture and connectivity [23].

While a protein–protein interaction can be presented by an edge between two participants, biochemical reactions and transport processes require different format of graphic representation. Number of participants of biochemical reaction (input and output reactants, and enzymes) is typically more than two, and biochemical pathway most frequently cannot be shown as a sequential series of pairwise interactions. An auxiliary central node was introduced for biochemical reactions to link an enzyme with input and output reactants. Such central node denotes the process itself and has no components. Similar approach was used for transport processes to link input, output molecular participants, and a transporter. Thus, auxiliary central nodes enable presenting pathways as logical and uninterrupted scheme of molecular process.

The direction of reactions could vary in the source pathway databases: "left to right", "right to left", "reversible". We converted each reaction or interaction in the format "left to right", and every reversible reaction was transformed into two coupled reactions.

In OncoboxPD, each gene product is named according to HGNC nomenclature [27]. Full names of non-protein molecules were saved in the PathBank database format, e.g. "Adenosine triphosphate". Additionally, molecular IDs are assigned, where possible, according to four different chemical databases: CAS [28], ChEBI [29], HMDB [30], PathWiz [31], and DrugBank [32]. If a molecule has no associated HGNC gene name, then it is not involved in calculation of pathway activation level (PAL) using gene expression data. However, such molecule may play a technically important role by connecting an overall pathway graph which is a prerequisite for algorithmic assignment of ARRs.

Table 3
Type and composition of pathway graph nodes in OncoboxPD.

Characteristic	Node type in OncoboxPD	Gene composition	Involvement in evaluation of ARR / PAL[a]	Example
Node with one or more gene products/components (proteins, nucleic acid molecules, small molecules, protein complexes, bound complexes)	participant	one or several gene products/components;	+/+ if node contains gene product	participant node *Ub* in HIF1Alpha pathway; gene products *UBB, UBC, UBD*
Node with name of biological effect, or with another crosslinking molecular pathway (entire pathway as single item)	participant	empty	+/-	participant node *Glycolysis* in 2-Ketoglutarate Dehydrogenase Complex Deficiency pathway
Auxiliary transport node	transport	empty	+/-	transport node *Ornithine* in Arginine and Proline Metabolism pathway. Citrulline transport from cytoplasm to mitochondrial matrix
Auxiliary biochemical reaction node	biochemical reaction	empty	+/-	reaction node *Glucose 1-phosphate + Uridine diphosphategalactose -> Galactose 1-phosphate + Uridine diphosphate glucose* in Congenital Disorder of Glycosylation CDG-IId pathway

[a] "+" and "-" indicates involvement of node in evaluation of ARR coefficients or PAL values.

3.3. Functional annotation of molecular pathways

In OncoboxPD, an attempt was made to functionally characterize all pathways according to their molecular physiological implication. To this end, we used Gene Ontology (GO) annotations [33], and analyzed gene sets corresponding to molecular pathway components using *enrichGO* function of ClusterProfiler R package [34] to identify biological processes which are statistically significantly linked with each individual pathway. We then assigned these specific GO tags to the respective molecular pathways and functional groups of pathways were, therefore, formed as those having common GO tag(s) (Fig. 1). Altogether, we identified 6 485 such functional groups (Fig. 1).

Different functional groups differed ~ three orders of magnitude by their representation and included 1–1021 molecular pathways (Fig. 2). The biggest functional groups with more than 800 pathways are listed in Table 4.

3.4. Pathway activation level calculation

Pathway activation level (PAL) is a metric allowing direct calculation of pathway activity levels using high-throughput gene expression data e.g. obtained by transcriptomic or proteomic screens. It can take positive and negative values in case of up- and downregulation of a pathway, respectively [1,7,35]. PAL values can serve to characterize molecular processes in-depth and in a large scale [19,35], or can be used as the biomarkers for many aspects of human biology including molecular pathology and personalized medicine [36–38]. In this version of OncoboxPD, we include PAL calculation and visualization tool that can supplement the database with the functional analysis of the pathways.

OncoboxPD has Web-based built-in tool for PAL calculation that is available at https://open.oncobox.com. User can upload expression data of interest to interrogate pathway activation compared to controls. PALs are calculated according to [35] as follows:

$$PAL_p = 100 * \sum_n ARR_{n,p} * \lg(CNR_n) / \sum_n |ARR_{n,p}|,$$

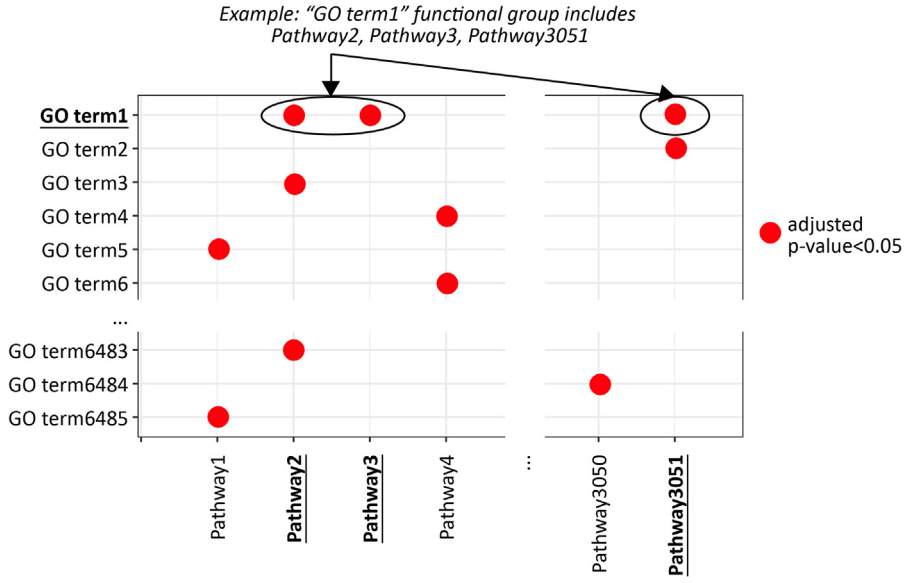

Fig. 1. Schematic representation of molecular pathway functional classification according to GO terms enrichment. Each functional group corresponds to a specific GO term and includes pathways, where this GO term is statistically significantly enriched (adjusted *p*-value less than 0.05). In this assay, only the pathways with unique gene compositions were considered.

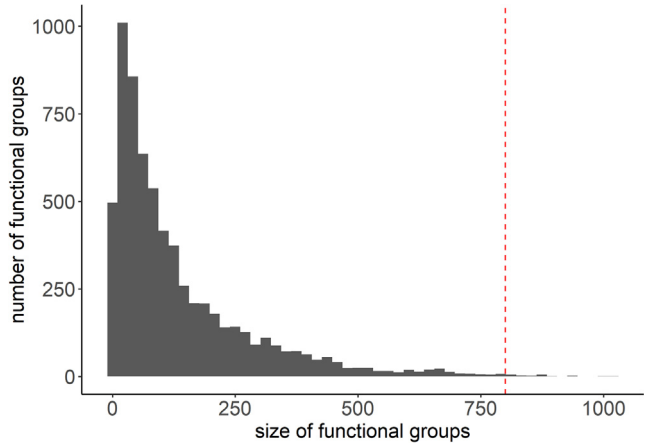

Fig. 2. Size distribution of GO functional groups of pathways (number of pathways included). Groups with more than 800 pathways are shown right to dashed red line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Functional OncoboxPD pathway groups with more than 800 members.

Pathway functional group ID (GO Tag)	Number of pathways included
Activation of protein kinase activity	1021
Peptidyl-serine phosphorylation	937
Fc receptor signaling pathway	896
Response to peptide hormone	883
Regulation of MAP kinase activity	882
Immune response-activating cell surface receptor signaling pathway	869
Neuronal death	858
Regulation of neuron death	834
Positive regulation of cellular protein localization	833
Blood coagulation	817
Gland development	810
Positive regulation of cell adhesion	802

where PAL_p is PAL for pathway p, CNR_n is case-to-normal ratio, ratio of gene n expression level in a sample under study to average level in control group; ARR (see above; activator/repressor role) is a Boolean value that depends on function of gene *n* product in pathway *p*. ARR can take values of − 1 when n inhibits p; 1 when n activates p; 0 when n has ambiguous or unknown role in *p*; 0.5 and − 0.5, when n is rather activator or inhibitor of p, respectively.

PAL calculator can use as the controls custom data provided by the user, or (default setting) human tissue RNAseq profiles obtained from healthy donors killed in road accidents. User can upload file with expression data and create a new analysis. At this stage, user can select pathway database, and can use an auxiliary option "scores for control samples" that enables calculating PAL values for the control samples to assess variations in the control group. The output results are returned as ready-to-download tables with CNR and PAL values for all samples under analysis. Alternatively, the results can be explored, analyzed and visualized with the web service. By clicking on the results, user can obtain overall PAL table, Sample table and pathway activation chart for the group of samples under analysis. Clicking on Sample table returns separate PAL datasheets for each sample. For every pathway in a given sample, its pathway activation chart can be visualized by the software. The sample-specific datasheet also includes CNR for all genes under analysis (for example, for up to 36,183 genes when profiles from the default collection are used as the norms), and log2(CNR) values. In every sample profile, each pathway ID can be clicked to obtain tabs for differentially expressed genes included in this pathway and their CNRs; pathway nodes and components; static and dynamic pathway graphic scheme.

Pathway activation chart is automatically generated for every sample or every group of samples (Fig. 3). It summarizes overall PAL calculation results and returns top 10 most strongly activated pathways (ordered top to bottom) and top 10 most strongly inhibited pathways (ordered bottom to top) with PAL values, their *p*-values and FDR-adjusted *p*-values (Benjamini – Hochberg correction). The principle of *p*-value calculation is shown on Table 5.

Pathway activation chart represents top most strongly activated and most strongly inhibited molecular pathways in a sample/group of samples. The number of displayed top pathways can be selected by the user. When a group of samples is investigated, then *t*-test *p*-value is shown for the comparison between case and control groups. All graphic materials generated are available for download as.*png* and.*svg* files. The detailed guidelines for PAL calculation and analysis with step-by-step screenshots is available through open.oncobox.com. Also, three thyroid cancer expression profiles[39] are preloaded for each user as demo example of samples with PAL calculating results (using six healthy thyroid samples from ANTE collection as control group).

Alternatively, we also developed a Python library *oncoboxlib,* that can be run on local computer and can be freely downloaded by the user. To calculate PAL values, *Oncoboxlib* requires files with HGNC gene symbols [27] and the corresponding expression levels for at least one case and at least one control sample. Installation instruction for *Oncoboxlib* and demo-example are available in Supplementary File 1 and at https://pypi.org/project/oncoboxlib.

3.5. Pathway visualization.

OncoboxPD collection is supplemented by pathway visualization tool. It allows to interactively visualize pathway structure and internal molecular interactions in the format of static or dynamic directed graph (Fig. 4 A, B).

Therein, static graph automatically obtains optimal layout to avoid node or line overlap and to adapt node size to its labels (Fig. 4A). As an option, user can switch from optimal to compact layer mode to decrease distances between the nodes. In turn, dynamic graph is interactive scheme that can be moved using Cytoscape plugin, where user can customize the layer by drugging the items (nodes) with mouse. This option is helpful in case of complicated interactions and long node labels.

Color of each node reflects logarithm of mean CNR for all of its components and corresponds to a scale given below with green color for upregulated, and red color for downregulated nodes. The default scale is designed to represent all lg(CNR) variabilities present on the graph for the case under investigation. However, user can customize color intensity by manually selecting scale limits, but this can lead to equal maximum color intensity for some nodes if they exceed the scale threshold. Such nodes will be marked by bold black frame. The nodes shown by grey color have no gene components and no CNR values. Unaffected or poorly affected nodes (with CNR $\sim$ 1, and lg(CNR) $\sim$ 0) are shown white.

Red arrows show inhibitory interactions, green arrows - activating interactions; arrows for ambiguous or poorly investigated interactions are shown black. Arrows involved in biochemical reactions or transport processes are comprised as the activating interactions. Auxiliary central nodes of biochemical reactions and transport processes have rhombic shape and are filled in black with no label because they denote the process but not its molecular participants (Fig. 5). Such nodes without gene products involved have no CNR values, but they are necessary for ARR assignment to all pathway graph members.

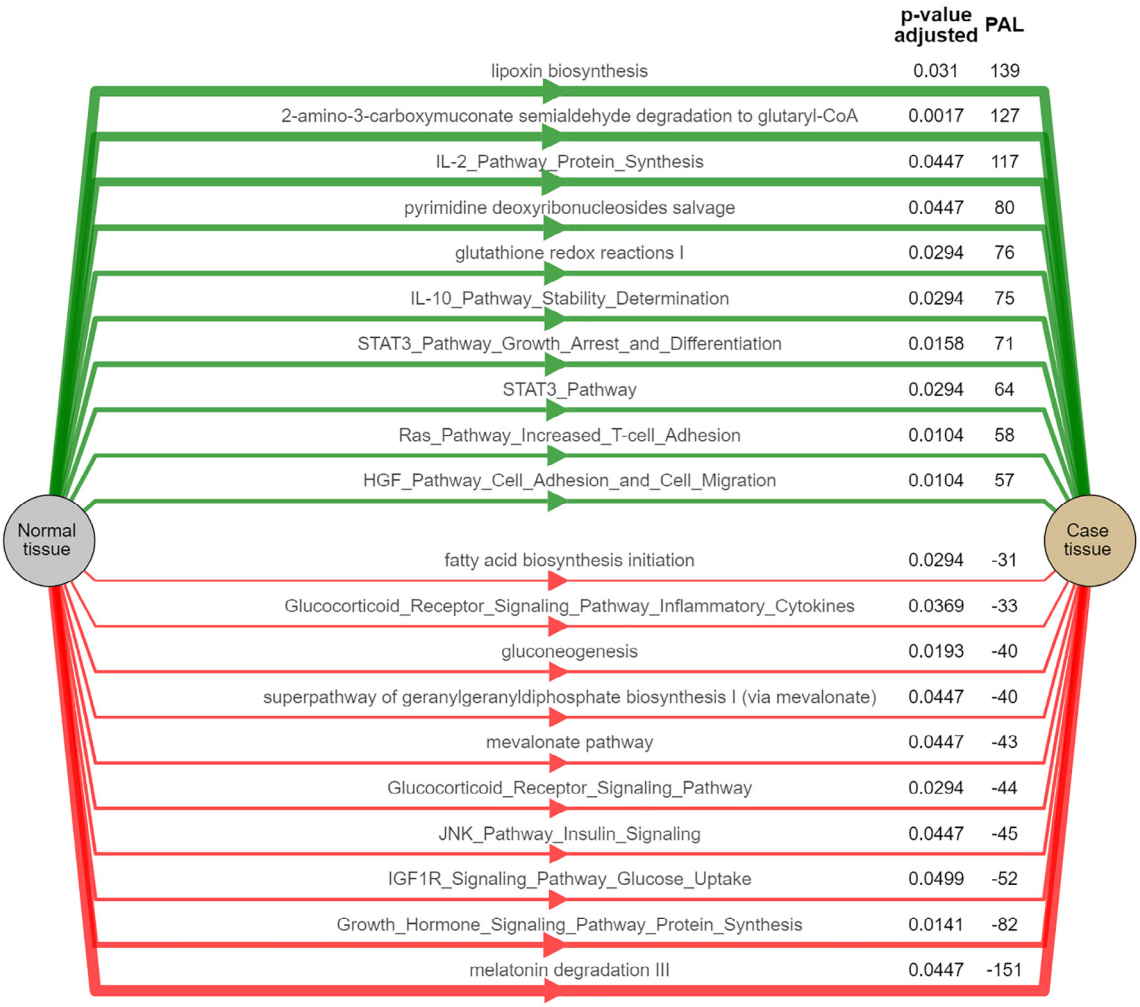

Fig. 3. Example of pathway activation chart. Green lines show top 10 most strongly activated pathways (ordered top to bottom), red lines show top 10 most strongly inhibited pathways (ordered bottom to top). Thickness is proportionate to absolute value of PAL. In this example, RNA sequencing gene expression profiles of three thyroid cancer samples [39] were compared with six healthy thyroid normal samples from ANTE collection [40]. PAL values, *t*-test *p*-values and FDR-adjusted p-values are shown right to the pathway names. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Principle of *p*-value calculation for pathway activation chart.

Number of control samples	Number of case samples (replicates) under analysis	Method of *p*-value calculation
3 or more	3 or more	*t*-test
3 or more	1	*p*-value is defined as a quantile of PAL in a sample investigated relatively to PAL distribution in control samples
less than 3	any number	*p*-value is not calculated

3.6. Human interactome graph of protein interactions and metabolic reactions

Using collection of pathways as the knowledgebase of molecular interactions, we also built combined human interactome and metabolome model. This is the directed graph, where nodes are genes or metabolites, and edges are known pairwise molecular interactions present in the OncoboxPD. The model was visualized using Gephi software and ForceAtlas2 algorithm [41] (Fig. 6).

Totally, we used molecular architectures of 50 178 different pathways. Complex pathway nodes containing *n* molecular participants were divided into *n* nodes with only one participant. Thus, each vertex represents one pathway participant on the graph. We then combined all pathway graphs together based on the coinciding gene products and metabolites. The resulting interactome graph includes 122 929 vertices and 600 137 edges. It incorporates totally 64 095 molecular participants and 361 654 interactions (excluding auxiliary nodes and interactions).

From all these pathways, we excluded molecular participants which were not connected within the overall network (less than 1% of the initial pathway members). The remaining molecular interactors formed a connected graph. The graph density was $8.08*10^{-5}$, average vertex degree (the number of edges connecting the vertex) was 4.9. However, some vertices had extremely high vertex degree, for example, 25 336 for Cytidine monophosphate, 24 034 for Coenzyme A, 23 300 for gene product *CRLS1*, 22 718 for gene product *DGAT1*, 2 813 for gene product *PEMT*, and 2 799 for S-Adenosylmethionine. The interactome model built enables finding the shortest path between genes of interest, or to identify gene interactomic neighborhood (e.g. at the distance of one, two or three edges). The model built is available in *graphml* format (https://doi.org/10.6084/m9.figshare.16617676.).

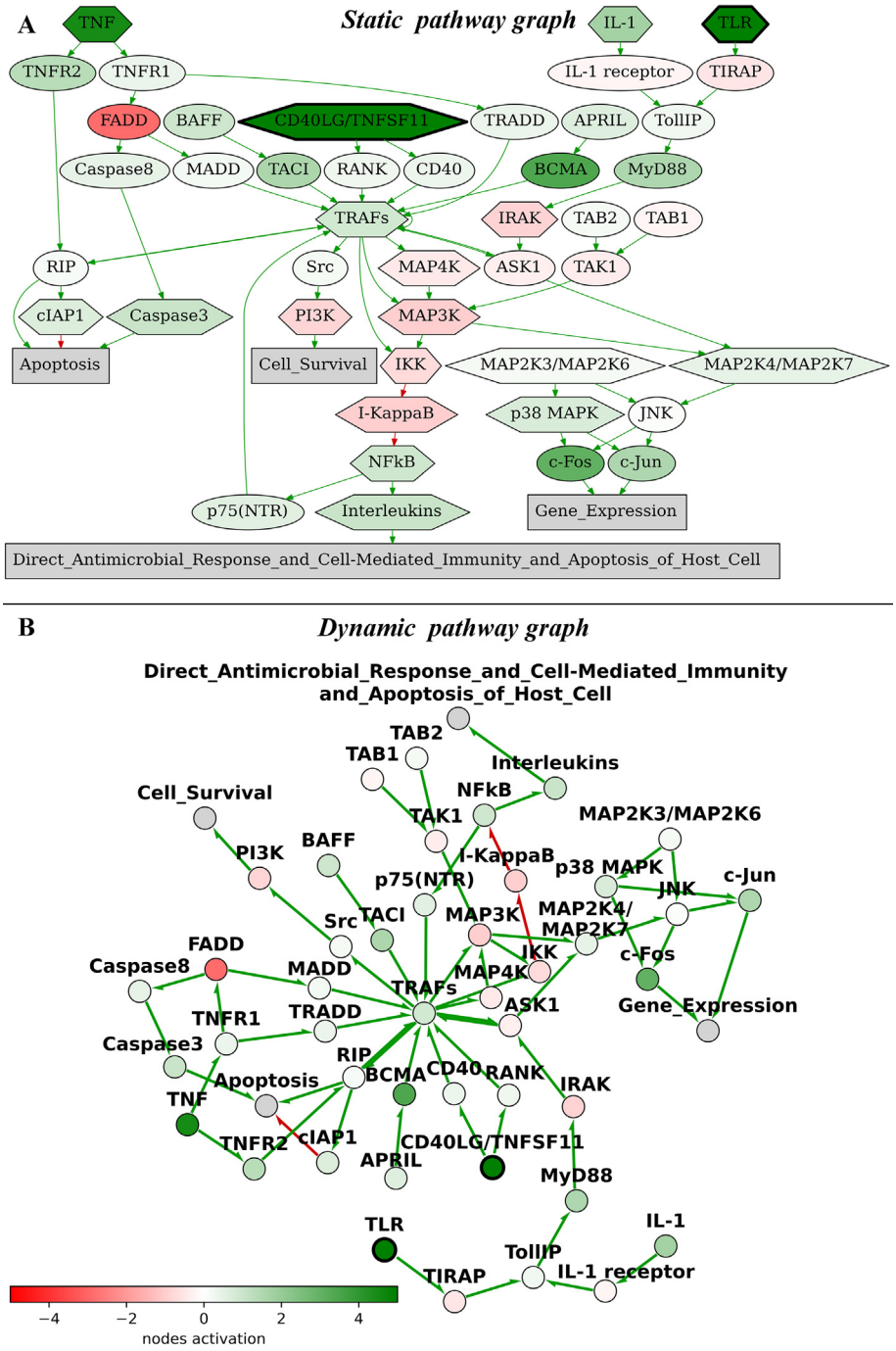

Fig. 4. TRAF molecular pathway visualization using OncoboxPD software. Nodes correspond to individual pathway components or to their complexes. Color of every node reflects logarithm of mean CNR for all node components, according to a scale given with green upregulated, and red downregulated nodes. Grey nodes have no gene components, and no CNR values. Nodes that are unaffected (with CNR ~ 1) are shown white. Red and green arrows stand for inhibitory and activating interactions, respectively. Ovals denote gene products, octagons - metabolites, hexagons – complexes, and other nodes (text labels without recognized molecular components) are shown as rectangles. A bold black border indicates outlier values that are outside of scale limits. In this example, RNA sequencing gene expression profile of thyroid papillary cancer sample TC15 [39] was compared against six healthy thyroid samples from ANTE collection [40]. A) Static pathway graph. B) Projection of dynamic interactive pathway graph. The figure can be found following the path: result file in Folders("example" with green icon)->sample TC15 in Sample table-> Pathway activation level tab->TRAF Pathway-> static and dynamic graphs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. 3.7.Comparison with the previous pathway aggregator databanks

There are several previously published knowledge bases that systematically aggregate information related to molecular pathways. For example, OmniPath accumulates data in the form of five databases including annotations of signaling network interactions, enzyme-post-translational modification relationships, characteristics of individual proteins, protein complexes, and of their roles in intercellular communication [42,43]. OmniPath has web-based, R, Cytoscape, and python applications which can generate and visualize custom (e.g. tissue specific) molecular networks. However, there is no option of calculating functional molecular pathway activation metrics using gene/protein expression data.

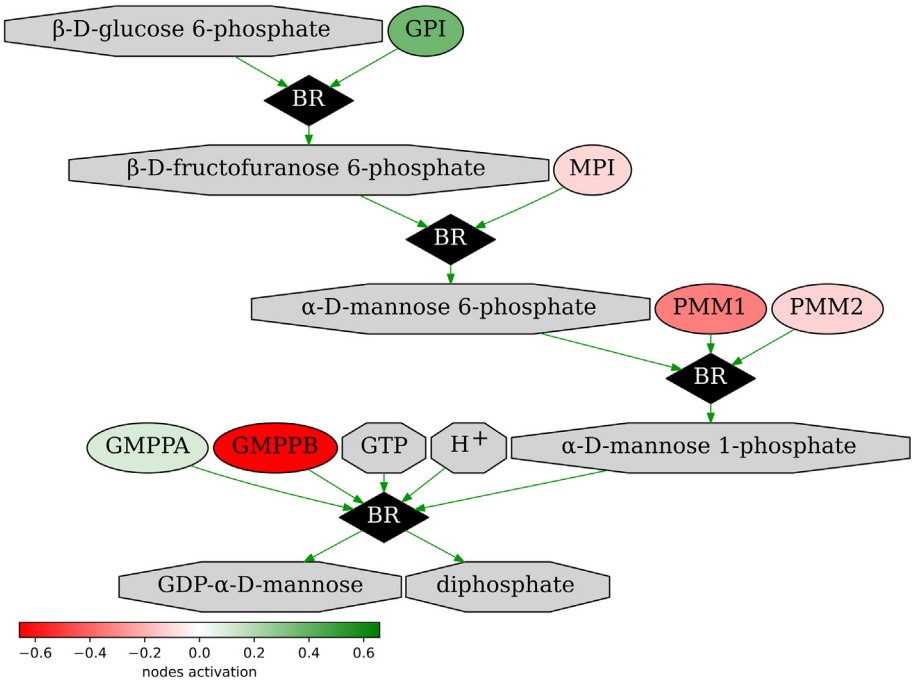

Fig. 5. OncoboxPD visualization of *GDP-mannose biosynthesis* molecular pathway. Nodes correspond to pathway participants. Color reflects node activation according to color scale on the bottom. For this example, thyroid papillary cancer sample 5 RNA sequencing profile [39] was normalized on six healthy thyroid samples from ANTE collection [40]. Nodes that don't correspond to known gene products are shown grey because for them no CNR value can be calculated. Auxiliary central nodes of biochemical reactions (BR) have rhombic shape and are filled in black. Ovals denote gene products, octagons - metabolites. Green arrows denote activating interactions. The figure can be found following the path: result file in Folders ("example" with green icon) ->sample TC15 in Sample table-> Pathway activation level tab-> De novo triacylglycerol biosynthesis pathway -> static graph. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In turn, Scalable Precision Medicine Open Knowledge Engine (SPOKE) resource combines several relevant databases of not only molecular pathways, but also of diseases, symptoms, biological processes, genes, relevant drugs and their side effects [44]. It can generate and visualize molecular networks based on the built-in databases. However, the source datasets cannot be extracted by the user.

In ConsensusPathDB database, individual molecular interactions are catalogued, and user can interrogate enrichment of specific gene networks [45]. However, the pathways are available for download as the non-organized lists of genes or metabolites, without information on their molecular interactions. Instead, all molecular interactions are available as a single dataset without link to pathways and with hidden type of interaction. However, for a fraction of pathways and interactions this information can be viewed via built-in online visualization tool, but in every case the interaction type is not annotated with the functional effect(s) on a pathway, such as *activation/inhibition/other* in the OncoboxPD.

The Pathway Commons tool combines many pathways-related databases, and has R, Cytoscape, and Java packages enabling to visualize and to certain extent to analyze the pathways [46]. On the other hand, it cannot calculate pathway activation metrics, and has no option of building pathway activation charts.

PathMe [47] database uniformly merged three large primary sources of molecular pathways (KEGG, Reactome, WikiPathways). PathMe transforms different original interaction and node types to Biological Expression Language (BEL) types to assess similarity of the same pathways from different sources (e.g., mTOR pathway from KEGG, Reactome, WikiPathways). Such similar pathways then can be merged as the larger networks according to degree of a coincidence of the same molecular participants and interactions.

We conclude, therefore, that OncoboxPD has the following advantages:

-presenting molecular pathways in a uniform format that is ready for directly functionally assessing pathway activation metrics;

-built-in tool for calculating pathway activation levels (PALs) for 50 K + molecular pathways using custom RNA/protein expression data;

-quick graph overview of most strongly differentially regulated pathways;

-visualization of pathway activation charts in two alternative (static and dynamic) modes, where color of each node of a given pathway reflects its up/downregulation in a sample of interest;

However, OncoboxPD has the following limitations:

-a user cannot create custom pathways in a web-tool;

-similar pathways from different database are given separately and not merged;

-molecular interactants are not attributed by cell/tissue localization;

-only molecular pathways are listed, without separate pairwise molecular interactions;

–no local application for pathway visualization is available to download.

Main distinguishing features of OncoboxPD in comparison with other pathway aggregator databanks is shown in Table 6.

5. Discussion

We report here OncoboxPD collection that includes 51 672 molecular pathways, which is to our knowledge currently the biggest human molecular pathways database with functionally characterized individual pathway nodes. Furthermore, all pathways are functionally classified according to GO terms enrichment patterns. OncoboxPD is a structured curated collection of protein-

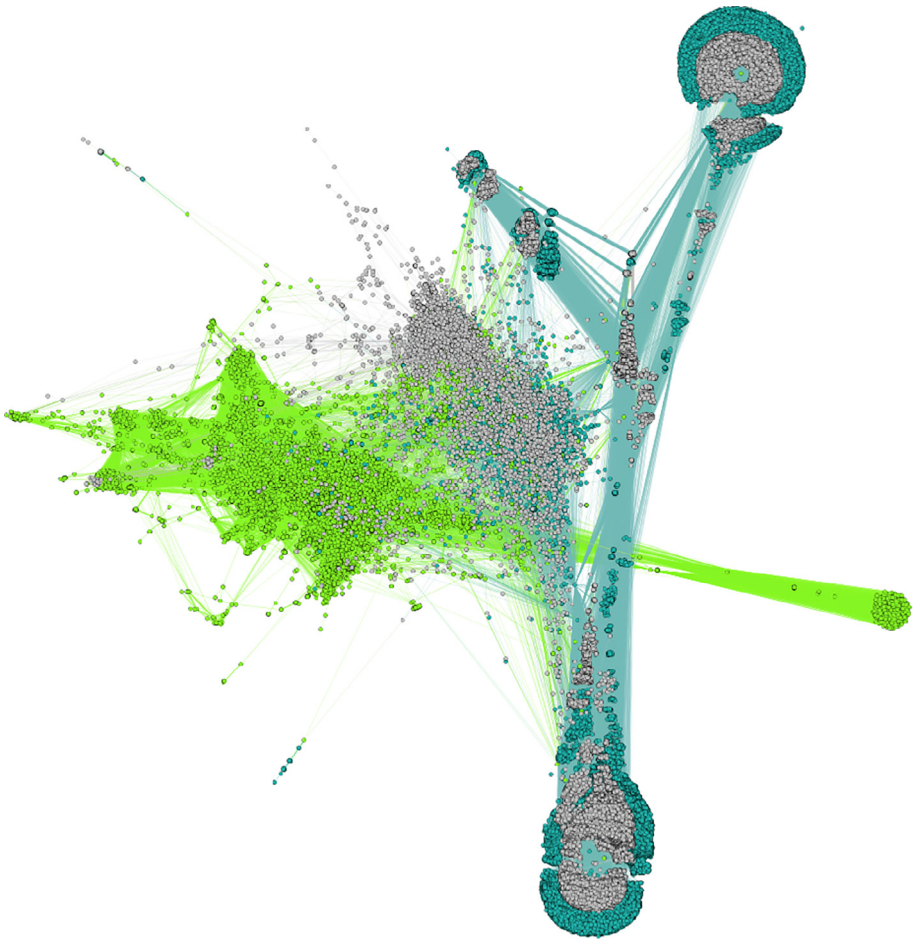

Fig. 6. Human interactome model of protein interactions and metabolic reactions. Graph vertices represent pathway participants: gene products (green), metabolites (blue) and auxiliary nodes/nodes with label of biological effect (grey). Graph edges are interactions between pathway participants. Edges inherit color from donor nodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

based and of metabolic human molecular pathways. All pathway participants, their interactions and reactions are uniformly processed and annotated, and are ready for numeric analysis of experimental expression data.

We did not aggregate here neither databases which may contain data with remarkably different levels of evidence like WikiPathways that can be edited by the users, nor datasets of pairwise molecular interactions (like IntAct). Unfortunately, we couldn't quantitatively estimate the reliability of each piece of data included in OncoboxPD because pathway construction was done by expert teams managing the source databases using different approaches, datasets, and algorithms. However, we can highlight some technical limitations of data presentation and storage which are peculiar to certain source knowledgebases. First, using of non-machine-readable formats may lead to loss of certain information. For example, KEGG database uses graphical.png format and original xml-based KGML format, and ceased to support universal BioPAX format. In KEGG, graphical format frequently contains information that is not included in machine-readable KGML format, e.g. biological process names, conditions of interactions, labels of spatial compartments, or separately given participants without clear links to other interactors.

In turn, PathBank database provides several alternative formats for each pathway: BioPAX, SBGN, SBML, PWML, RXN, and SDF. But only one of them, i.e. PWML (PathBank original format), contains all the information from the corresponding visual map. However, even there some elements may have only coordinates for the picture, but no information about their link with other process participants. Such labels are absent in other formats, thus leading to partial loss of data.

Furthermore, the universal format BioPAX has no clear direction and type of interaction in "subpathway" items in several cases, when interactors are simultaneously "activators" and "activated" or "inhibitors" and "inhibited" without a clear link between certain role and certain interaction or reaction. Qiagen SABiosciences database provided data only in the graphic format that had to be manually curated during construction of OncoboxPD. Another important issue is maintenance of the above source databases, their regular updates and availability. For example, the databases Biocarta and NCI PID ceased to exist in their original form, but were saved in various aggregator resources after specific data processing, that theoretically could alter data completeness.

In the examples given, we explained PAL calculation for RNA expression data. However, the same algorithms may be also applied for the quantitative proteomic profiles. In addition, algorithmic assessment of pathway mutation enrichment, e.g. in cancer samples [48–50], will be another possible direction of developing this database. The current pathway activity assessment interface is tailored to analize gene expression/proteomic data, but further updates may increase its functionality to assess also high-throughput metabolomic profiles. Another direction of future investigations is cross-linking gene/protein expression profiles with metabolomic data for the hybrid pathways including both gene products and metabolites, as shown on Fig. 5. Such integra-

Table 6
Major characteristics of selected pathway aggregating databanks.

Database (DB) name	OncoboxPD	Pathway Commons [46]	ConsensusPathDB [45]	SPOKE [44]	OmniPath [42,43]	PathMe [47]
Number of DBs/ pathway DBs	7/7	22/8	31/12	47/4	103/11	3/3
Human pathway databases included	Biocarta, HumanCyc, KEGG, NCI, PathBank, Reactome, Qiagen	Reactome, NCI, HumanCyc, PANTHER, KEGG, INOH, NetPath, Pathbank	Reactome, KEGG, Humancyc, NCI, Biocarta, Netpath, INOH, Ehmn, Pharmgkb, Smpdb, Signalink, Wikipathways	NCI and Reactome from Common pathways, KEGG (number of pathways is not available), WikiPathways	AlzPathway, Ma'ayan 2005, CancerCellMap/ NetPath, CST, Macrophage, KEGG, NCI, PANTHER, Reactome, SPIKE, WikiPathways	Reactome, KEGG, Wikipathways
Number of pathways	51,672	5,772	5,578	1,822	Not annotated	3095
Number of interactions	391,327	2,424,055	864,683	2,250,197	507,997	greater than215,000
Non-pathway interactions	–	+	+	+	+	–
Participant type	molecular	molecular	molecular	11 types (molecular and and others)	molecular	molecular
Web built-in visualization tool for:	canonical pathways	canonical pathways, interactions	canonical pathways, their fragments, interactions	interactions, custom pathways	Not available	merged canonical pathways, interactions
Local applications	Python	R, Java, Cytoscape	Cytoscape	–	Python, R, Cytoscape	Python
Functional analysis by pathway classification	+	–	–	–	–	–
Intracellular localization of pathway participants	–	+	–	–	–	–
Web analysis of custom gene expression data	Scoring of pathway activation	–	Enrichment/over-representation analysis	–	–	–
Construction of custom pathways	–	–	+	+	+	–
Annotated effect of interactions (activation, inhibition, neutral)	+	–	–	–	+	+

tion requires differential weighting of the PAL values obtained after gene/protein expression data and metabolomic profiles, which is currently unsolved problem and will be a matter of further studies.

Five out of seven source databases of OncoboxPD collection contain only protein–protein interactions (Table 1), and two remaining datasets contain also metabolite interaction data. To our knowledge, there is still no analytic system for pathway-level processing of high-throughput metabolomic data, and this may be important direction of future OncoboxPD development. For example, quantitative metabolomic data could be employed to calculate activation levels for the metabolic pathways, where the case-to-normal ratios (CNRs) will be found by comparing metabolite concentrations in the case and control biosamples.

In the future, we plan to maintain OncoboxPD a growing database that will be regularly updated when new pathways, or their more relevant functional annotations become available.

In OncoboxPD, we for the first time uniformly algorithmically classified large collection of molecular pathways according to the physiological processes they are involved by using Gene Ontology terms. The same approach may be employed for annotating new pathway collections, and such functional labels may be important to identify groups of relevant processes specific to certain conditions. For example, it was shown recently that signaling, cytoskeleton, metabolic and DNA repair pathways have distinct features in mutation accumulation and transcriptional activation in cancer [19].

Finally, OncoboxPD is currently the collection of human pathways and related analytic tools, whereas in the future it can be expanded to a number of model organisms where large-scale interactomic data are available.

Availability of data and materials.

OncoboxPD database and the corresponding built-in tools are freely available at https://open.oncobox.com. OncoboxPD can be accessed in two ways: by starting a new session (no registration needed), or by continuing previous session (authorization required). More detailed description of the access modes is given in the Help section at open.oncobox.com. The option of PAL calculation is also possible using *oncoboxlib* Python library, available at https://gitlab.com/oncobox/oncoboxlib.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

MZ, VT, AGa, AGu, MSo have a financial relationship with Omicsway Corp. The remaining authors declare that they have no conflict of interest.

Acknowledgments

We thank OmicsWay research initiative for technological support, software and access to source pathway databases. Cloud-based computational facilities for this study were supported by Amazon and Microsoft Azure grants.

Funding

This work was supported by Russian Science Foundation grant (21-74-20066). YW was supported by the National Natural Science Foundation of China grant (No. 81800805), Qingdao Key Research Project (No. 17-3-3-10-nsh and 19-6-1-3-nsh), and Qingdao Key Health Discipline Development Fund.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.006.

References

[1] Aliper AM, Korzinkin MB, Kuzmina NB, Zenin AA, Venkova LS, Smirnov PY, et al. Mathematical Justification of Expression-Based Pathway Activation Scoring (PAS). Methods Mol Biol 2017;1613:31–51. https://doi.org/10.1007/978-1-4939-7027-8_3.
[2] Borisov N, Aksamitiene E, Kiyatkin A, Legewie S, Berkhout J, Maiwald T, et al. Systems-level interactions between insulin-EGF networks amplify mitogenic signaling. Mol Syst Biol 2009;5. 10.1038/msb.2009.19.
[3] Kholodenko BN, Demin OV, Moehren G, Hoek JB. Quantification of short term signaling by the epidermal growth factor receptor. J Biol Chem 1999;274:30169–81. https://doi.org/10.1074/jbc.274.42.30169.
[4] Kiyatkin A, Aksamitiene E, Markevich NI, Borisov NM, Hoek JB, Kholodenko BN. Scaffolding protein Grb2-associated binder 1 sustains epidermal growth factor-induced mitogenic and survival signaling by multiple positive feedback loops. J Biol Chem 2006;281:19925–38. https://doi.org/10.1074/jbc.M600482200.
[5] Chowdhury S, Sarkar RR. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. Database 2015;2015. https://doi.org/10.1093/database/bau126.
[6] Wishart DS, Li C, Marcu A, Badran H, Pon A, Budinski Z, et al. PathBank: A comprehensive pathway database for model organisms. Nucleic Acids Res 2020;48:D470–8. https://doi.org/10.1093/nar/gkz861.
[7] Buzdin A, Sorokin M, Garazha A, Sekacheva M, Kim E, Zhukov N, et al. Molecular pathway activation - New type of biomarkers for tumor morphology and personalized selection of target drugs. Semin Cancer Biol 2018;53:110–24. https://doi.org/10.1016/j.semcancer.2018.06.003.
[8] Chartier M, Najmanovich R. Detection of Binding Site Molecular Interaction Field Similarities. J Chem Inf Model 2015;55:1600–15. https://doi.org/10.1021/acs.jcim.5b00333.
[9] Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. Int J Proteomics 2014;2014:1–12. https://doi.org/10.1155/2014/147648.
[10] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. Nucleic Acids Res 2009;37:D674–9. https://doi.org/10.1093/nar/gkn653.
[11] QIAGEN - Pathway-Central n.d. https://www.qiagen.com/us/shop/genes-and-pathways/pathway-central/ (accessed September 19, 2018).
[12] Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio - The analysis and navigation of molecular networks. Bioinformatics 2003;19:2155–7. https://doi.org/10.1093/bioinformatics/btg290.
[13] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res 2016;44:D481–7.
[14] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
[15] Elkon R, Vesterman R, Amit N, Ulitsky I, Zohar I, Weisz M, et al. SPIKE - A database, visualization and analysis tool of cellular signaling pathways. BMC Bioinf 2008;9. https://doi.org/10.1186/1471-2105-9-110.
[16] Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al. The MetaCyc database of metabolic pathways and enzymes-a 2019 update. Nucleic Acids Res 2020;48:D455–D1453. https://doi.org/10.1093/nar/gkz862.
[17] Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 2004;6:R2. https://doi.org/10.1186/gb-2004-6-1-r2.
[18] von Eichborn J, Dunkel M, Gohlke BO, Preissner SC, Hoffmann MF, Bauer JMJ, et al. SynSysNet: integration of experimental data on synaptic protein–protein interactions with drug-target relations. Nucleic Acids Res 2012;41:D834–40. https://doi.org/10.1093/nar/gks1040.
[19] Zolotovskaia MA, Tkachev VS, Seryakov AP, Kuzmin D V., Kamashev DE, Sorokin MI, et al. Mutation enrichment and transcriptomic activation signatures of 419 molecular pathways in cancer. Cancers (Basel) 2020;12. 10.3390/cancers12020271.
[20] Villaveces JM, Koti P, Habermann BH. Tools for visualization and analysis of molecular networks, pathways, and -omics data. Adv Appl Bioinforma Chem 2015;8:11–22. https://doi.org/10.2147/AABC.S63534.
[21] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 2012;92:414–7. https://doi.org/10.1038/clpt.2012.96.
[22] Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol 2011;696:291–303. https://doi.org/10.1007/978-1-60761-987-1_18.
[23] Buzdin AA, Sorokin M, Borisov NM, Kuzmin D, Gudkov A, Zolotovskaia MA, et al. Algorithmic annotation of functional roles for components of 3044 human molecular pathways. Front Genet 2021;12:139. https://doi.org/10.3389/FGENE.2021.617059.
[24] Nishimura D. BioCarta Biotech Softw Internet Rep 2001;2:117–20. https://doi.org/10.1089/152791601750294344.
[25] Nakaya A, Katayama T, Itoh M, Hiranuka K, Kawashima S, Moriya Y, et al. KEGG OC: A large-scale automatic construction of taxonomy-based ortholog clusters. Nucleic Acids Res 2013;41:D353–7. https://doi.org/10.1093/nar/gks1239.
[26] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2014;42:D472–7. https://doi.org/10.1093/nar/gkt1102.
[27] Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA, et al. The HGNC and VGNC resources in 2017. Nucleic Acids Res 2017;45:D619–25. https://doi.org/10.1093/nar/gkw1033.
[28] Huffenberger MA, Wigington RL. CHEMICAL ABSTRACTS SERVICE APPROACH TO MANAGEMENT OF LARGE DATA BASES. J Chem Inf Comput Sci 1975;15:43–7. https://doi.org/10.1021/ci60001a013.
[29] de Matos P, Alcántara R, Dekker A, Ennis M, Hastings J, Haug K, et al. Chemical entities of biological interest: An update. Nucleic Acids Res 2009;38. https://doi.org/10.1093/nar/gkp886.
[30] Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 2018;46:D608–17. https://doi.org/10.1093/nar/gkx1089.
[31] Pon A, Jewison T, Su Y, Liang Y, Knox C, Maciejewski A, et al. Pathways with PathWhiz. Nucleic Acids Res 2015;43:W552–9. https://doi.org/10.1093/nar/gkv399.
[32] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014;42:D1091–7. https://doi.org/10.1093/nar/gkt1068.
[33] Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis SE, et al. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47:D330–8. https://doi.org/10.1093/nar/gky1055.
[34] Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi A J Integr Biol 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.
[35] Borisov N, Sorokin M, Garazha A, Buzdin A. Quantitation of Molecular Pathway Activation Using RNA Sequencing Data. Methods Mol Biol 2020;2063:189–206. https://doi.org/10.1007/978-1-0716-0138-9_15.
[36] Sorokin M, Ignatev K, Poddubskaya E, Vladimirova U, Gaifullin N, Lantsov D, et al. RNA sequencing in comparison to immunohistochemistry for measuring cancer biomarkers in breast cancer and lung cancer specimens. Biomedicines 2020;8. https://doi.org/10.3390/BIOMEDICINES8050114.
[37] Poddubskaya EV, Baranova MP, Allina DO, Smirnov PY, Albert EA, Kirilchev AP, et al. Personalized prescription of tyrosine kinase inhibitors in unresectable metastatic cholangiocarcinoma. Exp Hematol Oncol 2018;7:21. https://doi.org/10.1186/s40164-018-0113-x.
[38] Poddubskaya E, Bondarenko A, Boroda A, Zotova E, Glusker A, Sletina S, et al. Transcriptomics-Guided Personalized Prescription of Targeted Therapeutics for Metastatic ALK-Positive Lung Cancer Case Following Recurrence on ALK Inhibitors. Front Oncol 2019;9:1026. https://doi.org/10.3389/fonc.2019.01026.
[39] Vladimirova U, Rumiantsev P, Zolotovskaia M, Albert E, Abrosimov A, Slashchuk K, et al. DNA repair pathway activation features in follicular and papillary thyroid tumors, interrogated using 95 experimental RNA sequencing profiles. Heliyon 2021;7:. https://doi.org/10.1016/j.heliyon.2021.e06408e06408.
[40] Suntsova M, Gaifullin N, Allina D, Reshetun A, Li X, Mendeleeva L, et al. Atlas of RNA sequencing profiles for normal human tissues. Sci Data 2019;6:36. https://doi.org/10.1038/s41597-019-0043-4.
[41] Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 2014;9:. https://doi.org/10.1371/journal.pone.0098679e98679.
[42] Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods 2016;13:966–7. https://doi.org/10.1038/NMETH.4077.
[43] Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol 2021;17. 10.15252/MSB.20209923.
[44] Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful

machine-readable embeddings. Nat Commun 2019.1–10.;2019(101):10. https://doi.org/10.1038/s41467-019-11069-0.
[45] Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB–a database for integrating human functional interaction networks. Nucleic Acids Res 2009;37. https://doi.org/10.1093/NAR/GKN698.
[46] Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res 2011;39. https://doi.org/10.1093/NAR/GKQ1039.
[47] Domingo-Fernández D, Mubeen S, Marín-Llaó J, Hoyt CT, Hofmann-Apitius M. PathMe: Merging and exploring mechanistic pathway knowledge. BMC Bioinf 2019;20:1–12. https://doi.org/10.1186/S12859-019-2863-9/FIGURES/5.
[48] Zolotovskaia M, Sorokin M, Garazha A, Borisov N, Buzdin A. Molecular Pathway Analysis of Mutation Data for Biomarkers Discovery and Scoring of Target Cancer Drugs. Methods Mol Biol 2020;2063:207–34. https://doi.org/10.1007/978-1-0716-0138-9_16.
[49] Zolotovskaia MA, Sorokin MI, Emelianova AA, Borisov NM, Kuzmin DV, Borger P, et al. Pathway based analysis of mutation data is efficient for scoring target cancer drugs. Front Pharmacol 2019;9. https://doi.org/10.3389/fphar.2019.00001.
[50] Zolotovskaia MA, Sorokin MI, Roumiantsev SA, Borisov NM, Buzdin AA. Pathway instability is an effective new mutation-based type of cancer biomarkers. Front. Oncol 2019;9. https://doi.org/10.3389/fonc.2018.00658.