Title

Metabox: A Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration

Permalink

https://escholarship.org/uc/item/7v84k1v9

Journal

PLOS ONE, 12(1)

ISSN

1932-6203

Authors

Wanichthanarak, Kwanjeera
Fan, Sili
Grapov, Dmitry
et al.

Publication Date

2017

DOI

10.1371/journal.pone.0171046

Peer reviewed

RESEARCH ARTICLE

Metabox: A Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration

Kwanjeera Wanichthanarak[1☯], Sili Fan[1☯], Dmitry Grapov[1], Dinesh Kumar Barupal[1], Oliver Fiehn[1,2]*

[1] West Coast Metabolomics Center, Genome Center, University of California Davis, Davis, California, United States of America, [2] Biochemistry Department, King Abdulaziz University, Jeddah, Saudi Arabia

☯ These authors contributed equally to this work.
* ofiehn@ucdavis.edu

Abstract

Similar to genomic and proteomic platforms, metabolomic data acquisition and analysis is becoming a routine approach for investigating biological systems. However, computational approaches for metabolomic data analysis and integration are still maturing. Metabox is a bioinformatics toolbox for deep phenotyping analytics that combines data processing, statistical analysis, functional analysis and integrative exploration of metabolomic data within proteomic and transcriptomic contexts. With the number of options provided in each analysis module, it also supports data analysis of other 'omic' families. The toolbox is an R-based web application, and it is freely available at http://kwanjeeraw.github.io/metabox/ under the GPL-3 license.

OPEN ACCESS

Citation: Wanichthanarak K, Fan S, Grapov D, Barupal DK, Fiehn O (2017) Metabox: A Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration. PLoS ONE 12(1): e0171046. doi:10.1371/journal.pone.0171046

Editor: Matej Orešič, Steno Diabetes Center, DENMARK

Received: October 14, 2016

Accepted: January 13, 2017

Published: January 31, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by U24 National Institutes of Health and National Institute of Diabetes and Digestive and Kidney Diseases DK097154. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist. Majority of Dr.

Introduction

Advances in high-dimensional 'omic' platforms have enabled large-scale characterization of molecular phenotypes for a variety of biological systems. Processing, integrating and visualizing modern biological data sets is a formidable task that requires a flexible computational framework to enable a growing variety of genomic, biochemical and phenotypic data types. Key metabolomic data analyses are comprised of four major steps (1) raw data pre-processing including compound identification (2) data processing including data transformation and data normalization (3) statistical analysis and (4) data interpretation. Raw data pre-processing composes of several steps to preprocess raw signals from analytical techniques (e.g. mass spectrometry (MS) and nuclear magnetic resonance (NMR)), which includes noise reduction, peak picking and compound identification [1]. Metabox is independent of tools that were used for raw data preprocessing. Metabox starts with step (2), data processing. Data transformation is defined as the process of converting data into more useful forms of the same data either by mathematical operations (e.g. log-transformation) or by changing formats (e.g. rounding data, or converting units). Data normalization and data transformation are performed on the matrix of imported metabolomic result data to minimize systematic and technical variations before

Dmitry Grapov's contribution to the manuscript occurred while he was not part of CDS and instead the Principal Statistician at the University of California Davis. We have changed his affiliation to reflect this fact and he maintains his stated no conflict of interest.

statistical evaluation [1] Subsequently, statistical analyses are used to pinpoint metabolites that are altered based on the experimental design group that therefore must be detailed in the imported data. Statistcal result outputs are then interrogated by downstream modules including network- and pathway-based data analyses and visualization tools [2]. Fig 1 illustrates key analysis steps that can be performed in metabox.

Given the diverse areas of metabolomic data analysis, computational platforms are needed for average users to increase efficiency and to reduce hurdles for in-depth interpretations of metabolomic data along with other 'omic' data types. MetaboAnalyst is an online web application that covers comprehensive analysis of metabolomic data [3–5]. The functional module for the integrative analysis is based on metabolomics pathway analysis (MetPA). In difference to MetaboAnalyst, metabox uses a graph database as underlying resource. Pathway relationships in MetaBox are dynamically constructed and therefore, pathway mapping is not limited to static (pre-defined) metabolic pathways as in MetaboAnalyst. [6]. Similarly, a tool like IMPaLa addresses the integration of transcriptomic and metabolomic data on the metabolic pathway contexts but it lacks of an integrative visualization option [7]. In fact, metabolite profiles incorporate environmental and genetic factors into biochemical processes [8]. Combining metabolic information with biological mechanisms of genes and proteins reflects the flow of biological information between layers of cellular regulation. An integrative exploration of metabolomics with secondary lines of molecular 'omic' information enables generating new hypotheses such as identification of disease factors [9]. Further tools exist that also exploit the framework of biological networks for multi-omic data exploration, such as MetScape [10], Grinn [2], ConsensusPathDB [11] or 3Omics [12]. Similar to metabox, these tools facilitate the joint visualization of genes, proteins and metabolites in different combinations of biological networks. However, in general these programs are less powerful in data processing and statistical analysis tools. In addition, both MetScape and Grinn lack modules for pathway analyses. All these programs serve the community as open-access software; commercial software such as MetaCore was not available to us for comparison.

We here introduce the R-based web application 'metabox' that combines state-of-the-art methods essential for metabolomic data processing, statistical analyses, network-based visualization and functional analyses. The current version of metabox highlights deep analyses of metabolomic data and integrative exploration of metabolites, proteins and genes of interest within several contexts of biological networks such as metabolic pathways, gene regulation and molecular interaction. A number of options are provided, which can be selected to apply on the other omic studies (e.g. transcriptomics and proteomics). Briefly, the following features are integrated in metabox: (i) data normalization and data transformation, (ii) univariate statistical analyses with hypothesis testing procedures that are automatically selected based on users' study designs, (iii) joint visualization of genes, proteins and metabolites in different combinations of biochemical networks, (iv) calculation of data-driven networks using correlation-based approaches, (v) estimating chemical structure similarity networks from substructure fingerprints, (vi) functional interpretation with overrepresentation analysis, functional class scoring and WordCloud generation and (vii) interactive visualization of information-rich tables and networks. In addition, metabox is distributed as a standard R package which allows ease of distribution, installation and makes internal metabox functions available to be used in custom workflows by advanced users.

Materials and Methods

Metabox is an R-based software package that is developed as a web application for interactive scientific computing and visualization. The tool is composed of a pre-compiled graph database and R functions essential for the analysis of metabolomic data. The following sections explain each of the tool components in details.

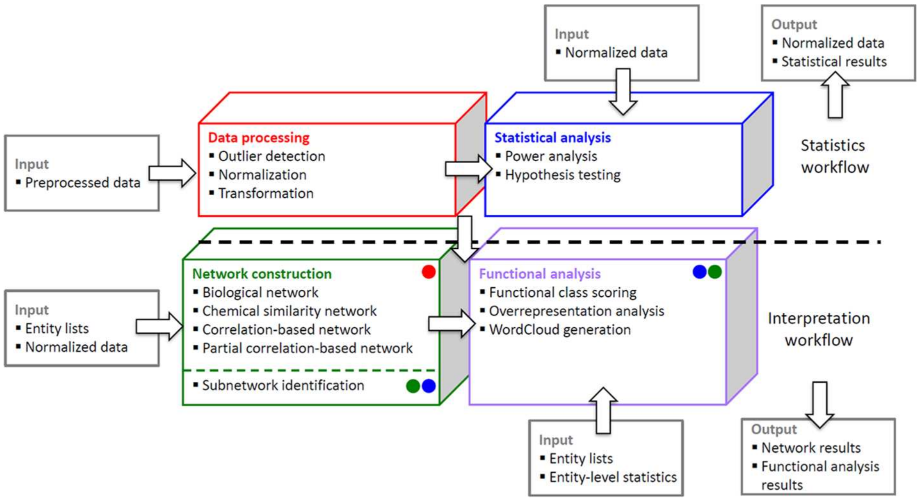

Fig 1. Analysis workflows. Metabox supports in-depth analysis of metabolomic data by including four analysis modules: data normalization (red), statistical analysis (blue), network construction (green) and functional analysis (purple). Outputs from each module are in red, blue, green and purple circles respectively. The tool accepts external inputs on each analysis level. Within metabox, the output from an analysis module can be used for subsequent analyses in the other modules denoted as a colored circle inside a box.

doi:10.1371/journal.pone.0171046.g001

Internal graph database

The internal graph database is a pre-compiled graph database, which is used for collecting prior knowledge from several resources (S1 Table). The database is required for biological network query functions explained in the following section. A Neo4j graph database (http://neo4j.com/) is used for storing information of molecular entities and their relationships. In this version, the graph database is available for *Homo sapiens*. Upon installation, it will automatically access this pre-compiled human database on our server.

The database schema was adapted from the Grinn package [2] (S1 Fig). In particular, the database includes the majority of biological relationships among molecular entities such as substrate-product relationships (BIOCHEMICAL_REACTION), protein-compound catalysis (CATALYSIS), protein-protein interactions (MOLECULAR_BINDING), gene-encoding proteins (CONVERSION), protein-gene transcription regulations (CONTROL), gene-gene associations (GENETIC_ASSOCIATION), microRNA-gene regulations (CONTROL) and pathway-entity annotations (ANNOTATION) (S2 Table).

Two identifier systems are used in the database: Neo4j internal identifiers (NIDs) and Grinn identifiers (GIDs). The NID is a numeric number generated automatically by the Neo4j database system. The GID uses the authentic identifiers of different domain databases. In particular, GIDs uses ENSEMBL identifiers for genes, miRTarBase nomenclature for micro-RNAs, UniProt entries for proteins, and PubChem CIDs or KEGG numbers for compounds and pathways. User inputs require either GID identifiers or convert other identifiers into NIDs via a converter tool provided in metabox (S3 File).

R-based functions for analysis workflows

In metabox, metabolomics analysis workflows are divided into statistics and interpretation (Fig 1). The statistics package includes data processing and statistical analysis, while the interpretation workflow is used for biological interpretations of the statistical outputs. Backend functions for both workflows were developed in the R programming language (https://www.r-project.org/).

Data processing and statistical analysis. Data processing and statistical analysis are the key tasks of metabolomic data analysis. The aim of data processing (transformation and normalization) is to improve normality of data sets in order to improve comparability of metabolite intensities. Subsequently, statistical tools are used to find significant molecular entities, both for hypothesis testing (univariate analysis) and cluster analysis (multivariate analysis which may serve for hypothesis generation).

- *Data normalization*. Two types of normalization methods are currently implemented: feature-based normalization and sample-based normalization. Both normalizations can be performed sequentially. Sample-based normalization aims to normalize each sample to control for the systematic and technical variance between samples. Several commonly used normalization methods are included in metabox: normalization to sample-specific metadata (such as dry mass), normalization to the sum of all identified metabolites [13], normalization to batch-median values of specific samples data, or LOESS (locally estimated scatterplot smoothing) normalization to specifically introduced quality control or pool samples [14]. LOESS normalizations are performed with span parameters that are automatically selected by cross-validation and batch effect correction [15]. Conversely, metabolite-based normalization aims to make the measured variables more comparable to each other with respect to total variance. Three methods are available in metabox including auto-scaling, Pareto scaling and range scaling [16]. These normalization methods are usually performed for multivariate analyses tools in order to ensure that data heteroscedasticity is reduced to a minimum.
- *Data transformation*. Metabox offers common variable transformation methods including logarithm and power. These methods can be used to improve data normality assumptions for parametric statistical hypothesis testing procedures which can otherwise be sensitive to non-normal distributions, outliers and lack of homogeneity of variances.
- *Exploratory data analysis*. Metabox uses principal components analysis (PCA) score plots for real-time visualization during data processing procedures [16]. It allows users to detect outliers and choose appropriate methods for data normalization and transformation according to the properties of the data structure. In addition, users may select scatters on the PCA score plots and get the corresponding sample information from a donut chart. This helps users to discover unexpected features within the data structure.
- *Univariate Analysis*. Metabox collects a variety of well-established statistical hypothesis testing methods and post hoc analysis procedures (Table 1). In addition, metabox includes corresponding non-parametric testing procedures, post hoc analysis with false discovery rate (FDR) correction on both main effect level and simple main effect level, and power analyses at the entity-level. Bootstrapping is provided as an optional non-parametric testing procedure. Furthermore, metabox automatically and appropriately suggests statistical analysis methods according to the user-input study design. This feature aims to aid users through the depths of statistical terminology.
- *Power analysis* is provided at the entity-level. It covers power analyses for hypothesis testing procedures listed in Table 1. Metabox offers two levels of power analyses: prior to estimate sample size and post-hoc to estimate statistical power.

Network construction. Network-based analysis is a promising approach to explore molecular interactions. In addition, networks can be used as a scaffold for mapping experimental results to identify potential markers, altered pathways or active subgraphs. Metabox supports both

Table 1. List of statistical analysis procedures in metabox.

Study Design	Test	Methods	Following Procedures
Two-independent group	Statistically significant difference exists between means of two independent groups	Welch's t test, Mann-Whitney U test*, t test;	P-values adjusted for FDR correction using Bonjamini-Hochberg procedure [17]
Two-paired group	Mean difference between paired observations is statistically significant different from zero	Welch's t test on difference, Mann-Whitney U test on difference*, Welch's t test on difference;	P-values adjusted for FDR using Bonjamini-Hochberg procedure
Multiple-independent group	There are statistically significant differences between the means of three or more independent groups	Welch ANOVA; Kruskal-Wallis rank sum test*; ANOVA;	Parametric test followed by Games-Howell (or Tukey) post hoc test to compare all possible combinations of group differences; non-parametric test followed by pairwise comparisons using Dunn's procedure [18] with Bonferroni adjustment.
Multiple-paired group	There are statistically significant differences between the means of three or more levels of a within-subjects factor.	Repeated ANOVA; Friedman test*;	Parametric test result corrected by Greenhouse-Geisser procedure [19] for violation of sphericity and followed by Bonferroni post hoc test as suggested by Maxwell and Delaney [20]; non-parametric followed by pairwise comparisons using Wilcoxon signed-rank tests [21] were performed with Bonferroni correction
Two-way independent groups	There is a statistically significant interaction effect between two ways of independent groups	Two-way ANOVA, two-way ANOVA with robust estimation	Results followed by post hoc analysis on main effect level and simple main effect level corresponding to secondary study design structure
Two-way paired groups	There is a statistically significant interaction effect between two within-subject ways	Two-way repreated ANOVA	Results followed by post hoc analysis on main effect level and simple main effect level corresponding to secondary study design structure
Two-way mixed groups	There are differences between independent groups over time	Mixed ANOVA	Results followed by post hoc analysis on main effect level and simple main effect level corresponding to secondary study design structure

* Non-parametric test.

Underlined method denotes the default method in metabox

doi:10.1371/journal.pone.0171046.t001

construction of networks from domain knowledge relationships and from empirical relationships. There are four main functions to construct networks from different contexts.

- *Biological networks* are queried from our pre-compiled graph database that contains *a priori* relationships from several resources in SimpleNetwork and HeterogeneousNetwork analyses. SimpleNetwork analysis is used to query biological networks of one type of relationship (e.g. protein-protein interaction networks) whereas HeterogeneousNetwork analyses query biological networks containing one or combinations of relationship types (e.g. biochemical reaction networks containing substrate-product and compound-enzyme associations). Lists of entities and relationship patterns are illustrated in the database schema (S1 Fig).
- *Weighted correlation networks* are computed from experimental data using pairwise or partial correlation approaches. Pairwise correlation analysis computes associations between every pair of entities. Metabox includes Pearson, Spearman or Kendall correlation analyses based on the WGCNA package [22]. Partial correlation analysis is performed based on the qpgraph package that estimates associations between entities while controlling effects from other entities [23, 24].
- *Chemical structure similarity networks* are computed from PubChem substructure fingerprints by using Tanimoto chemical similarity matrix calculations [25, 26].
- *A high-scoring subnetwork* is enumerated from biological networks, weighted correlation networks or chemical structure similarity networks. This subnetwork identification is based on the

BioNet package that calculates scores of network nodes from p-value results of statistical analyses and follows by a heuristic search for identification of the high-scoring subnetwork [27, 28].

Functional analysis. Functional analysis is provided to aid biological interpretations of the results from statistical analysis, the network outputs, and user-uploaded list of molecular entities. The current version of metabox supports KEGG [29] pathway-based analysis for compounds, proteins and genes. Additionally, metabox uses Medical Subject Headings (MeSH) [30] for compound interpretations by querying the 'chemicals and drugs' category in the PubChem database [31]. Three different analysis options are included.

- *Functional class scoring [32]* or *set enrichment analysis* assesses the significance of annotation terms from entity-level statistics (e.g. p-values of entities calculated from statistical analysis). Metabox integrates commonly used methods of gene set enrichment analysis from the piano package, including [33], Fisher 's combined probability test [34], Stouffer's method [35], Reporter features [36, 37], Median and Mean. The Fisher's method computes each set-level statistics by combining log-transformation of entity-level statistics, while the Stouffer's method calculates from aggregating Z scores of entity-level p-values. The Reporter method is similar to the Stouffer's method with an additional procedure of background distribution correction. The Median and Mean method are straightforward (i.e. each set-level statistics is the median and mean of entity-level statistics respectively.). The piano package uses a permutation approach to evaluate the significance of entity set of an annotation term.
- *Overrepresentation analysis* is to identify overrepresented annotation terms of the given list of entities using hypergeometric test. For the networks containing multiple types of nodes, the Fisher's method is used to combine p-values from hypergeometric test.
- *WordCloud generation* is a simple, graphical presentation of annotation terms where the font size of a word corresponding to number of members. It provides a quick summary of annotation terms of the given list of entities without any statistical test. The function is based on the R package tm [38] and the R package wordcloud (https://cran.r-project.org/web/packages/wordcloud/).

Graphical user interface

The metabox graphical user interface (GUI) was implemented with HTML, JavaScript and CSS. The backend R functions were deployed on a web browser using the OpenCPU JavaScript library [39], which allows users to execute the R functions on a web browser such as Firefox, Chrome, Internet Explorer and Safari. The GUI is a two-column layout with a side navigation bar containing the list of different functions. On the right side, page contents are changed correspondingly to the selected function. Collapsible tree views of MeSH terms are drawn using the D3.js JavaScript library (https://d3js.org/). Word clouds are plotted using the wordcloud package (https://cran.r-project.org/web/packages/wordcloud/). The Cytoscape.js JavaScript library [40] is used to create interactive networks. Network navigation such as pan, zoom and select can be performed using a mouse or a touchpad. The resulting networks can be downloaded as PNG format and tab-delimited text files which can be used in other software.

Results and Discussions

In this section, the utility of metabox for metabolomic data analysis is demonstrated using two independent data sets that measured metabolomic and gene expression profiles of lung adenocarcinoma and adjacent non-malignant lung tissues [41, 42]. The example data sets, corresponding tutorials and metabox user manual are provided as supplemental files (S1, S2 and S3 Files, respectively).

Omic data sets

The metabolomic data contain 39 malignant and adjacent non-malignant lung tissue samples measured by gas chromatography time-of-flight mass spectrometry (GC-TOF-MS) and pre-processed by the BinBase database [43]. 462 compounds were measured, 171 of which were structurally identified and associated by PubChem CID. The data from the BinBase database were uploaded to metabox for log transformation before statistical analysis using paired t-test. After Benjamini and Hochberg FDR (BH-FDR) adjustment at 5% ($pFDR < 0.05$), there were 131 significant compounds between cancer and control tissues.

Gene expression profiles of 58 lung adenocarcinoma and adjacent non-malignant lung tissues were acquired using Illumina HumanWG-6 v3.0 expression BeadChip platform and downloaded from the Gene Expression Omnibus database [44] as transcriptomic data set GSE32863. Differential gene expression analysis comparing lung adenocarcinoma and adjacent non-malignant lung tissues was performed with GEO2R [44]. P-values were adjusted with Benjamini and Hochberg false discovery rate at 5% ($pFDR < 0.05$). Differential gene expression analysis comparing tumor and control tissues reported 171 out of 21,204 genes in which $pFDR < 0.05$ and $|log2FC| > 2$.

Using metabox for in-depth analysis of metabolomic data

The following metabox example shows a comprehensive analysis of metabolomic data, progressing through a variety of common analyses steps including: data processing, statistical analyses, and biochemical display of results within a variety of contexts including structural similarity, pathway and based on other biochemical relationships. The pre-processed metabolic profiles were imported to metabox for data processing and statistical analysis. Subsequently the output was transferred to the network construction part to calculate the chemical structure similarity network of the measured compounds (i.e. 171 compounds have associated PubChem CID). The default threshold was used with a correlation coefficient $r_{xy}> 0.7$. The resulting lists of pair-wise metabolic correlations are then displayed as network graph that can be enhanced by further empirical or annotation information. Here, we applied the Functional Class Scoring option to estimate significantly enriched pathways. We found significantly enriched pathways (p-value < 0.05) to include arginine and proline metabolism, arginine biosynthesis and pentose and glucuronate interconversions. Fig 2 shows a part of the resulting similarity network overlaid with enriched pathways. Amino acids, ornithine, citrulline and glutamine were metabolically linked through the pathway 'arginine biosynthesis'. In addition, ornithine was metabolically linked to proline, 5-aminopentanoic acid and glutamic acid through the pathway 'arginine and proline metabolism'.

Metabox supports integrative exploration of significant genes and compounds from lung cancer studies in biological network context

Next, we used the lists of significantly dysregulated genes ($pFDR < 0.05$ and $|log2FC| > 2$) and significantly different compounds ($pFDR < 0.05$) from comparisons between paired lung tumors and non-malignant tissues to construct the biological network using the internal graph relationship database. The resulting network was downloaded and enhanced further by mapping with statistical information using Cytoscape [45].

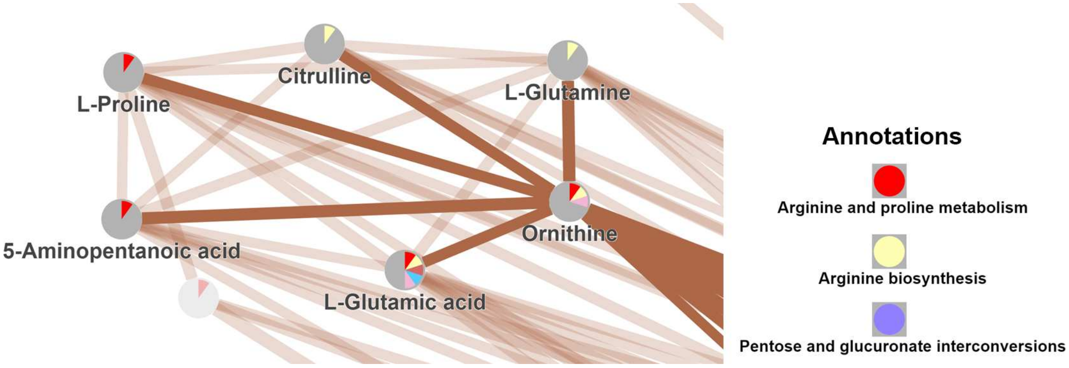

Fig 2. Partial visualization of the entire 'chemical structure similarity' network of metabolites in a lung adenocarcinoma study. Chemical similarities between all identified metabolites was calculated from PubChem substructure fingerprints. Network nodes are connected by correlation coefficients using edge thickess for correlations r_{xy}>0.7. Metabox functional class scoring was applied to estimate significantly enriched pathways (p<0.05), yielding arginine/proline metabolism, arginine biosynthesis and pentose/glucuronate interconversions among other pathways. Pathway enrichments are given by color in node pie charts.

doi:10.1371/journal.pone.0171046.g002

The output network shows metabolic links among carbohydrates, glycerol, fatty acids, palmitic acid, sphingolipids, sphinganine, amino acids, glutamic acid and glyceric acid, and glycosylation-related metabolites, including uridine diphosphate glucuronic acid (UDP-glucuronic acid) and uridine diphosphate-N-acetylglucosamine (UDP-GlcNAc) (Fig 3). Combinations of relationships can be explored through such networks, including protein-compound catalysis, protein-encoding genes and protein-gene regulations, integrating information from the regulation of transcription to cellular metabolism. For example, the network outlines metabolic relationships for UDP-GlcNAc and N-acetylglucosamine (GlcNAc) through the corresponding N-acetyltransferase enzymes encoded by *GCNT3* and *OGT* genes, both of which were significantly up-regulated. Modification of nuclear and cytosolic proteins by addition of O-linked β-N-actylglucosamine (O-GlcNAc) at serine or threonine residues or O-GlcNAcylation plays important roles in several cellular processes such as cell signaling, metabolism, transcription regulation and cell division [46]. Evidence indicates that hyper-O-GlcNAcylation is a common feature in cancer [46–48]. UDP-GlcNAc is a crucial GlcNAc donor, which is transferred to protein substrates by O-linked N-acetylglucosamine transferase [46]. Fig 3 also shows that the enzyme-encoding gene *OGT* was linked to protein regulators such as FOXO4 and TFE2 that were connected to genes of phospholipid phosphatase 3 (*PLPP3*) and long-chain-fatty-acid—CoA ligase 4 (*ACSL4*). Both genes were significantly down-regulated in lung cancer compared to control tissues and associated proteins were metabolically linked to significantly decreased compounds sphinganine and palmitic acid respectively. The *OGT* gene was also linked to transcription factors MYC (an oncoprotein found to involve in initiation and maintenance of several cancers [47, 49]) and LEF1, which in turn were associated to genes *ALDH18A1* and *GFPT1*. Both genes were significantly up-regulated in lung cancer compared to control samples and encoded proteins were metabolically linked to significantly increased glutamic acid. In addition, *ALDH18A1* shared the same transcription factor CREB1 with up-regulated gene *CHPF* of glucuronic acid transferase that was connected to substrate UDP-glucuronic acid.

This example demonstrates the use of metabox to integrate information from metabolic and gene expression results for joint visualization in the network context. The resulting network lists and shows relationships among protein regulators, genes, enzymes and metabolites related to amino acid, fatty acid and lipid metabolism, and glycosylation that can be functionally tested in subsequent studies.

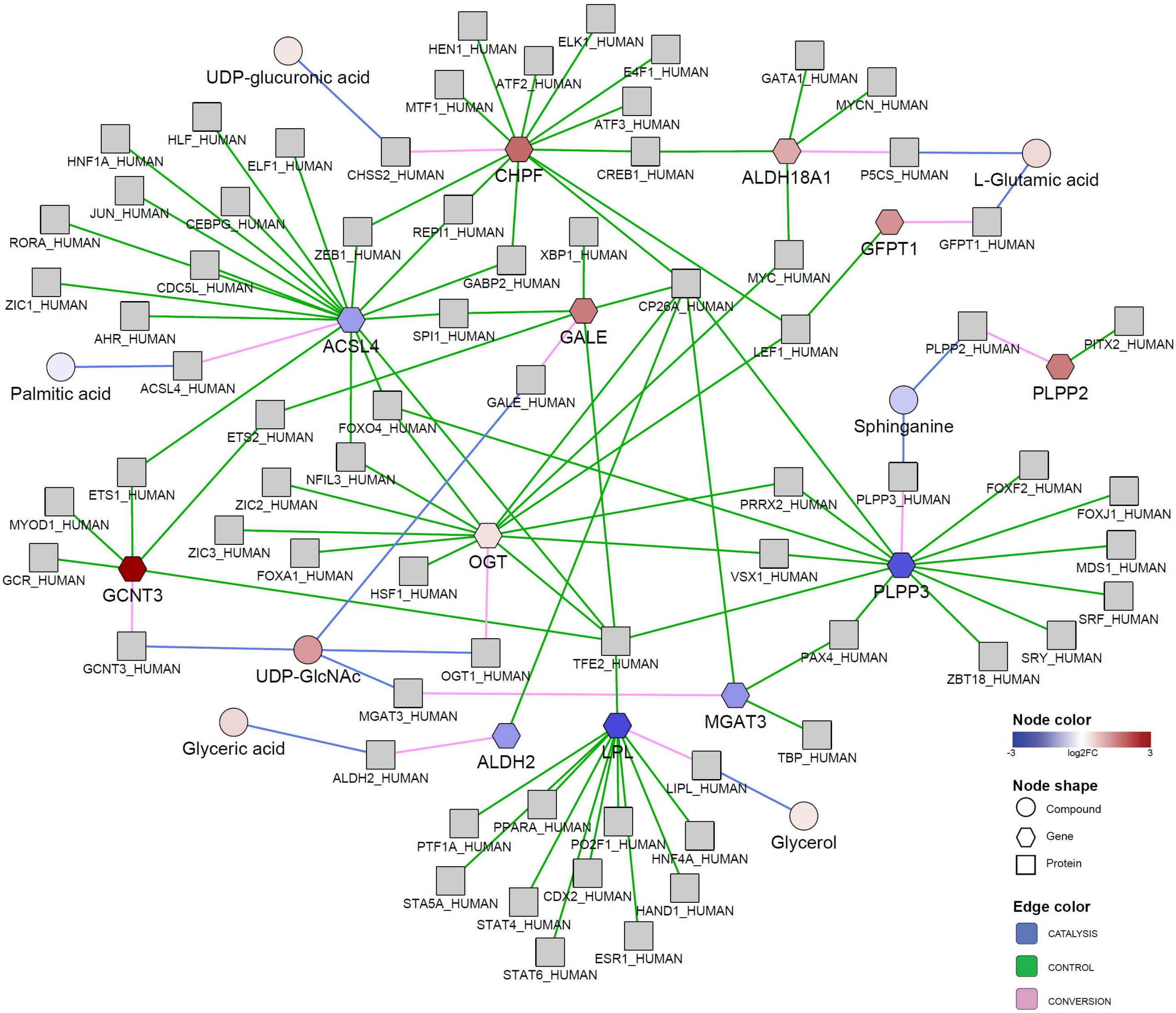

Fig 3. Biological network integrating significant differences in gene and metabolite regulation in lung adenocarcinoma compared to paired control tissues. Significantly different genes and metabolites were mapped onto the metabox internal graph database using enzymes as linking nodes (grey). The resulting network was downloaded and mapped relative changes between tumor and non-tumor tissues using Cytoscape. Graph relationships CONTROL, CONVERSION, and CATALYSIS are labeled by colored edges. The network shows metabolic links between glycerol, palmitic acid, sphinganine, glutamic acid, glyceric acid, UDP-glucuronic acid and UDP-GlcNAc.

doi:10.1371/journal.pone.0171046.g003

Comparison to existing tools

To date, a variety of bioinformatics tools exist for metabolomic data analysis, however, not all of them support the comprehensive analysis and integrative exploration of metabolomic data. Metabox is a freely available tool that consolidates a number of approaches for data processing, statistical analysis, network construction, integrative visualization and functional interpretations of metabolomic data in one software package. Table 2 provides an overview of tools that

Table 2. Comparison with existing tools for analysis and interpretation of metabolomic data.

Tools	Metabox	MetaboAnalyst	ConsensusPathDB	MetScape	3Omics	Grinn
Input	Gene lists, protein lists, compound lists, omic data, statistical values	Metabolomic data, mass spectral data, Zipped file of NMR data, mass spectral peak lists or mass spectral data	Gene lists, protein lists, compound lists, statistical values	Gene lists, compound lists, correlation values, statistical values	Omic data	Gene lists, protein lists, compound lists, omic data
Data processing	Yes	Yes	No	No	No	No
Statistical analysis	Univarate and power analysis, automatically suggest analysis method	Univariate and multivariate analysis, clustering, classification	No	No	No	No
Network construction	Biological-, weighted correlation- and chemical structure similarity network, subnetwork identification	No	Biological network	Biochemical network	Biochemical and weighted correlation network	Biochemical and weighted correlation network
Functional class scoring	Yes	Yes	Yes	No	No	No
Overrepresentation analysis	Yes	Yes	Yes	No	Yes	No
WordCloud generation	Yes	No	No	No	No	No
Output	Tab-delimited files for statistical results, network outputs, and functional analysis result, PNG file for network image, PDF, SVG or PNG file for WordCloud image	CSV files for analysis results, PDF, SVG, PNG or TIFF file for image	Tab-delimited files for network outputs, and functional analysis result	Cytoscape output files	Tab-delimited files for network outputs, and functional analysis result	Tab-delimited files for network outputs
GUI	Web-based	Web-based	Web-based	Cytoscape-based	Web-based	Web-based

doi:10.1371/journal.pone.0171046.t002

contains comparable features to metabox. It also highlights that metabox excels in the scope of statistics and integrative exploration with other 'omic' data in network contexts and functional analysis.

Limitations and future directions

The main aim of metabox is to support comprehensive analysis workflows of metabolomics and the integrative exploration with other 'omic' data. For the statistics workflow, we included standard normalization, transformation and statistical analysis approaches optimized for metabolomic data. While data processing approaches can be selectively applied on any 'omic' data sets, computational time is increased in larger data sets such as transcriptomics and proteomics. Furthermore, when using bootstrapping procedures, run times increase significantly. The current release of metabox does not yet provide a full list of all multivariate or advanced statistics tools (e.g. linear regression and Bayesian models) to support further statistical-based integration of multi-omic data sets. R-packages for these tools exist and will be integrated into the toolbox over time.

In the part of interpretation workflow, metabox supports joint visualization of molecular entities in several combinations of biological networks. The present databases include domain knowledge relationships available only for human from limited resources. We plan to update the current databases using more information from pre-compiled databases such as ConsensusPathDB. In addition, a function to combine different network outputs will also be added. Current functional interpretations are in the contexts of KEGG pathways and MeSH chemicals and drugs categories for compounds. We are investigating to include more types of biological annotations such as disease associations and gene ontology (GO) annotations [50], however, addition of further databases requires careful validation and suitability assessments.

Conclusions

In this study, we propose an alternative toolbox for thorough analysis and integrative exploration of metabolomic data. Metabox includes widely used statistical methods to process and identify keys entities of input experiments, offers different integrative analysis methodologies and provides interactive visualization to facilitate biological interpretations. The tool is embedded with a graph database and supports both knowledge-based and data-driven network construction. The design of GUI as an intuitive web interface aims to support bench biologists to simply perform data analysis. Metabox is also run as a standard R package for advanced users to use in combination with other R projects. The development of metabox highlights the needs of research communities for the efficient analysis, integration and interpretation of metabolomic studies.

Supporting Information

S1 Table. List of resources and type of information collected in the graph database.
(PDF)

S2 Table. List of nodes and relationships.
(PDF)

S1 Fig. Database schema. The database schema illustrates molecular entities and their relationships stored in the graph database. Ovals denote molecular entities and round rectangular indicate relationships.
(TIF)

S1 File. Example data sets.
(ZIP)

S2 File. Tutorials.
(PDF)

S3 File. User manual.
(PDF)

Acknowledgments

This project was funded by the National Institutes of Health, U24 DK097154 for the West Coast Metabolomics Center. All authors are thankful to Dr Johannes F. Fahrmann for valuable comments.

Author Contributions

Conceptualization: KW SF DG OF.

Data curation: KW SF.

Formal analysis: KW SF.

Funding acquisition: OF.

Investigation: OF.

Methodology: KW OF DB SF.

Project administration: OF DB.

Resources: OF.

Software: KW DG SF.

Supervision: OF DB.

Validation: KW OF DG DB SF.

Visualization: KW SF.

Writing – original draft: KW OF DB SF.

Writing – review & editing: KW OF DB SF DG.

References

1. Cambiaghi A, Ferrario M, Masseroli M. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Briefings in bioinformatics. 2016. Epub 2016/04/15.
2. Wanichthanarak K, Fahrmann JF, Grapov D. Genomic, Proteomic, and Metabolomic Data Integration Strategies. Biomarker insights. 2015; 10(Suppl 4):1–6. Epub 2015/09/24. PubMed Central PMCID: PMC4562606. doi: 10.4137/BMI.S29511 PMID: 26396492
3. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. Nucleic acids research. 2012; 40(Web Server issue):W127–33. Epub 2012/05/04. PubMed Central PMCID: PMC3394314. doi: 10.1093/nar/gks374 PMID: 22553367
4. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic acids research. 2009; 37(Web Server issue):W652–60. Epub 2009/05/12. PubMed Central PMCID: PMC2703878. doi: 10.1093/nar/gkp356 PMID: 19429898
5. Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic acids research. 2015. Epub 2015/04/22.
6. Xia J, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics. 2010; 26(18):2342–4. doi: 10.1093/bioinformatics/btq418 PMID: 20628077
7. Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics. 2011; 27(20):2917–8. Epub 2011/09/07. doi: 10.1093/bioinformatics/btr499 PMID: 21893519
8. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nature reviews Molecular cell biology. 2012; 13(4):263–9. Epub 2012/03/23. PubMed Central PMCID: PMC3682684. doi: 10.1038/nrm3314 PMID: 22436749
9. Pirhaji L, Milani P, Leidl M, Curran T, Avila-Pacheco J, Clish CB, et al. Revealing disease-associated pathways by network integration of untargeted metabolomics. Nature methods. 2016. Epub 2016/08/02.
10. Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics. 2012; 28(3):373–80. Epub 2011/12/03. PubMed Central PMCID: PMC3268237. doi: 10.1093/bioinformatics/btr661 PMID: 22135418
11. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic acids research. 2013; 41(Database issue):D793–800. Epub 2012/11/13. PubMed Central PMCID: PMC3531102. doi: 10.1093/nar/gks1055 PMID: 23143270
12. Kuo TC, Tian TF, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC systems biology. 2013; 7:64. Epub 2013/07/24. PubMed Central PMCID: PMC3723580. doi: 10.1186/1752-0509-7-64 PMID: 23875761

13. Borrego SL, Fahrmann J, Datta R, Stringari C, Grapov D, Zeller M, et al. Metabolic changes associated with methionine stress sensitivity in MDA-MB-468 breast cancer cells. Cancer & metabolism. 2016; 4:9. Epub 2016/05/04. PubMed Central PMCID: PMC4852440.

14. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nature protocols. 2011; 6(7):1060–83. doi: 10.1038/nprot.2011.335 PMID: 21720319

15. Wang SY, Kuo CH, Tseng YJ. Batch Normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration methods. Analytical chemistry. 2013; 85(2):1037–46. Epub 2012/12/18. doi: 10.1021/ac302877x PMID: 23240878

16. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC genomics. 2006; 7:142. Epub 2006/06/10. PubMed Central PMCID: PMC1534033. doi: 10.1186/1471-2164-7-142 PMID: 16762068

17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995; 57:289–300.

18. Dunn OJ. Multiple Comparisons Using Rank Sums. Technometrics. 1964; 6(3):241–&.

19. Greenhouse SW, Geisser S. On Methods in the Analysis of Profile Data. Psychometrika. 1959; 24 (2):95–112.

20. Rutherford A. Designing experiments and analyzing data: A model comparison perspective. Brit J Math Stat Psy. 2004; 57:185–7.

21. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bull. 1945; 1(6):80–3.

22. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9:559. Epub 2008/12/31. PubMed Central PMCID: PMC2631488. doi: 10.1186/1471-2105-9-559 PMID: 19114008

23. Tur I, Roverato A, Castelo R. Mapping eQTL networks with mixed graphical Markov models. Genetics. 2014; 198(4):1377–93. Epub 2014/10/02. PubMed Central PMCID: PMC4256758. doi: 10.1534/genetics.114.169573 PMID: 25271303

24. Castelo R, Roverato A. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. Journal of computational biology: a journal of computational molecular cell biology. 2009; 16 (2):213–27. Epub 2009/01/31.

25. Barupal DK, Haldiya PK, Wohlgemuth G, Kind T, Kothari SL, Pinkerton KE, et al. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC bioinformatics. 2012; 13:99. Epub 2012/05/18. PubMed Central PMCID: PMC3495401. doi: 10.1186/1471-2105-13-99 PMID: 22591066

26. Willett P. Chemical Similarity Searching. Journal of Chemical Information and Modeling. 1998; 38 (6):983–96.

27. Beisser D, Klau GW, Dandekar T, Muller T, Dittrich MT. BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics. 2010; 26(8):1129–30. Epub 2010/03/02. doi: 10.1093/bioinformatics/btq089 PMID: 20189939

28. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. 2008; 24(13):i223–31. Epub 2008/07/01. PubMed Central PMCID: PMC2718639. doi: 10.1093/bioinformatics/btn161 PMID: 18586718

29. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research. 2012; 40(Database issue):D109–14. Epub 2011/11/15. PubMed Central PMCID: PMC3245020. doi: 10.1093/nar/gkr988 PMID: 22080510

30. Lipscomb CE. Medical Subject Headings (MeSH). Bulletin of the Medical Library Association. 2000; 88 (3):265–6. Epub 2000/08/06. PubMed Central PMCID: PMC35238. PMID: 10928714

31. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic acids research. 2016; 44(D1):D1202–13. Epub 2015/09/25. PubMed Central PMCID: PMC4702940. doi: 10.1093/nar/gkv951 PMID: 26400175

32. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology. 2012; 8(2):e1002375. Epub 2012/03/03. PubMed Central PMCID: PMC3285573. doi: 10.1371/journal.pcbi.1002375 PMID: 22383865

33. Varemo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic acids

research. 2013; 41(8):4378–91. Epub 2013/02/28. PubMed Central PMCID: PMC3632109. doi: 10.1093/nar/gkt111 PMID: 23444143

34. Fisher RA. Statistical methods for research workers. Edinburgh, London,: Oliver and Boyd; 1925. ix p., 1 l., p.

35. Stouffer SA. The American soldier. Princeton,: Princeton University Press; 1949.

36. Oliveira A, Patil K, Nielsen J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC systems biology. 2008; 2(1):17.

37. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proceedings of the National Academy of Sciences. 2005; 102(8):2685–9.

38. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. Journal of statistical software. 2008; 25 (5):1–54.

39. Ooms J. The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns2014:[23 p.]. Available from: http://arxiv.org/abs/1406.4806.

40. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 2016; 32(2):309–11. Epub 2015/09/30. PubMed Central PMCID: PMC4708103. doi: 10.1093/bioinformatics/btv557 PMID: 26415722

41. Wikoff WR, Grapov D, Fahrmann JF, DeFelice B, Rom WN, Pass HI, et al. Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma. Cancer prevention research. 2015; 8 (5):410–8. Epub 2015/02/07. doi: 10.1158/1940-6207.CAPR-14-0329 PMID: 25657018

42. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. Genome research. 2012; 22(7):1197–211. Epub 2012/05/23. PubMed Central PMCID: PMC3396362. doi: 10.1101/gr.132662.111 PMID: 22613842

43. Fiehn O, Wohlgemuth G, Scholz M. Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. Lect Notes Comput Sc. 2005; 3615:224–39.

44. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2013; 41(Database issue):D991–5. Epub 2012/11/30. PubMed Central PMCID: PMC3531084. doi: 10.1093/nar/gks1193 PMID: 23193258

45. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003; 13 (11):2498–504. Epub 2003/11/05. PubMed Central PMCID: PMC403769. doi: 10.1101/gr.1239303 PMID: 14597658

46. Slawson C, Hart GW. O-GlcNAc signalling: implications for cancer cell biology. Nature reviews Cancer. 2011; 11(9):678–84. Epub 2011/08/19 PubMed Central PMCID: PMC3291174. doi: 10.1038/nrc3114 PMID: 21850036

47. Ma Z, Vosseller K. O-GlcNAc in cancer biology. Amino acids. 2013; 45(4):719–33. Epub 2013/07/10. doi: 10.1007/s00726-013-1543-8 PMID: 23836420

48. Hart GW, Slawson C, Ramirez-Correa G, Lagerlof O. Cross Talk Between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease. Annu Rev Biochem. 2011; 80:825–58. doi: 10.1146/annurev-biochem-060608-102511 PMID: 21391816

49. Dang CV. MYC on the path to cancer. Cell. 2012; 149(1):22–35. Epub 2012/04/03. PubMed Central PMCID: PMC3345192. doi: 10.1016/j.cell.2012.03.003 PMID: 22464321

50. Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic acids research. 2015; 43(Database issue):D1049–56. Epub 2014/11/28. PubMed Central PMCID: PMC4383973. doi: 10.1093/nar/gku1179 PMID: 25428369