~ Supporting Information ~

Ontolomics-P: ontology guide and data-driven analysis of mass spectrometry-based proteomics data

Yin Yang[1,4§], Shisheng Wang[1,2§], Wei Jiang[3], Youmei Jin[2], Wenjuan Zeng[2], Jingqiu Cheng[2], Dongbo Wu[3*], Hao Yang[2*] and Bairong Shen[4*]

[1] Department of Clinical Research Management, Proteomics-Metabolomics Core Facility, West China Hospital, Sichuan University, Chengdu, 610041, China

[2] Institutes for Systems Genetics and NHC Key Lab of Transplant Engineering and Immunology, West China Hospital, Sichuan University, Chengdu, 610041, China
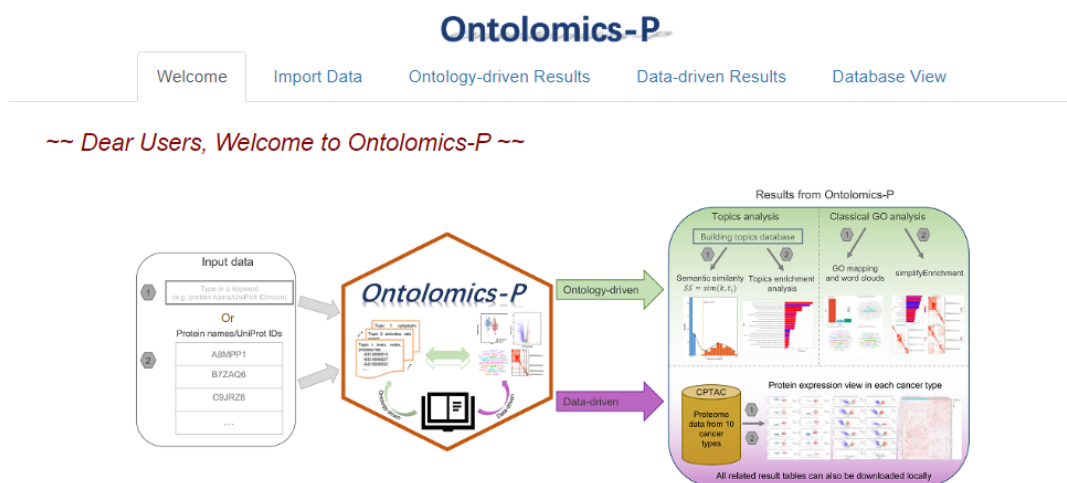
[3] Division of Infectious Diseases, State Key Laboratory of Biotherapy and Center of Infectious Disease, West China Hospital, Sichuan University, Chengdu 610041, China

[4] Joint Laboratory of Artificial Intelligence for Critical Care Medicine, Department of Critical Care Medicine and Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, 610041, China

Supplementary notes

Ontolomics-P stands out as a versatile tool tailored for the proteomics community, featuring a graphical user interface (GUI) to facilitate the analysis of MS-based proteomics data from ontology-driven and data-driven perspectives. This comprehensive tool incorporates four essential functions. Firstly, its Topic Analysis Integration employs natural language processing to categorize topics within a document collection, building a topics database based on the Gene Ontology (GO) database for efficient retrieval of protein/gene functional information. Secondly, *Ontolomics-P* implements topics enrichment analysis using Fisher's exact test method to identify significant associations and enrichments within the dataset. Thirdly, it includes classical GO enrichment analysis and semantic similarity analysis for diverse analytical approaches. Lastly, the tool seamlessly integrates quantitative proteomic data from ten cancer types stored in the CPTAC database, facilitating the efficient review of Profiles of Interest (POI) expression profiles in colon adenocarcinoma (COAD) [1], breast cancer (BRCA) [2], uterine corpus endometrial carcinoma (UCEC) [3], clear cell renal cell carcinoma (ccRCC) [4], lung adenocarcinoma (LUAD) [5], hepatocellular carcinoma (HECA) [6], head and neck squamous cell carcinoma (HNSCC) [7], pancreatic ductal adenocarcinoma (PADA) [8], lung squamous cell carcinoma (LSCC) [9], and ovarian cancer (OV) [10].

Users can visit this site: https://www.omicsolution.com/wukong/Ontolomics-P or https://www.omicsolution.org/wukong/Ontolomics-P. **Please Note:** If the online version does not work, which means you cannot open the links, it is probably because our server is down and we will fix it very soon. We also recommend users to install this tool locally (Please check "6. How to run this tool locally?" part below). Then the website homepage can be shown like this:

Ontolomics-P is a web-based tool, which possesses the core functions, including:

1. Topic analysis integration. Leveraging natural language processing, this feature allows researchers to identify and categorize various topics within a collection of documents. It builds a topics database based on the Gene Ontology (GO) database, aiding in the retrieval of pertinent protein/gene functional information;

2. Topics enrichment analysis implementation. Utilizing Fisher's exact test method, Ontolomics-P conducts topics enrichment analysis to identify significant associations and enrichments within the dataset;

3. Inclusion of classical GO enrichment analysis and semantic similarity analysis. To enhance practicality, Ontolomics-P retains classical GO enrichment analysis and semantic similarity analysis, offering diverse analytical approaches;

4. Integration of quantitative proteomic data from diverse cancer types. Ontolomics-P seamlessly integrates quantitative proteomic data from ten cancer types stored in the CPTAC database. These include colon adenocarcinoma (COAD), breast cancer (BRCA), uterine corpus endometrial carcinoma (UCEC), clear cell renal cell carcinoma (ccRCC), lung adenocarcinoma (LUAD), hepatocellular carcinoma (HECA), head and neck squamous cell carcinoma (HNSCC), pancreatic ductal adenocarcinoma (PADA), lung squamous cell carcinoma (LSCC), and ovarian cancer (OV).

In addition, this tool supports both online access and local installation. The source codes and installation instructions can be available in the GitHub repository: https://github.com/wangshisheng/Ontolomics-P under an MIT license.

Finally, Ontolomics-P is developed by R shiny (Version 1.6.0), and is free and open to all users with no login requirement. It can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. We would highly appreciate that if you could send your feedback about any bug or feature request to Shisheng Wang at shishengwang@wchscu.cn.

Friendly suggestions:

a) Open Ontolomics-P with Chrome, Mozilla Firefox, Safari or Firefox;

b) The minimum operating system specifications are: RAM 4GB, Hard drive 100 GB;

c) The monitor resolution (>= 1920x1080) is better.

^_^ *Enjoy yourself in Ontolomics-P* ^_^

1. Data Preparation

Users could either input a keyword (e.g., a protein name/UniProt ID/noun) or upload a list of proteins of interest (POIs).

1.1 A keyword (e.g., a protein name/UniProt ID/noun)

Herein, users could type in a protein name (e.g. TNFSF10), or a UniProt ID (e.g. P50591), or a noun (e.g. Liver).

1.2 A list of POIs

POIs (e.g. Differentially expressed proteins) could be saved in a .csv, .txt, .xlsx or .xls file as below:

	A
1	UniProtIDs
2	A0A096LP49
3	A8MPP1
4	B7ZAQ6
5	C9JRZ8
6	O00193
7	O00743
8	O14791
9	O15504
10	O43731
11	O60293
12	O60487
13	O60675
14	O60927
15	O75177

There is only one column in this file and each row is a protein name or UniProt ID. In addition, *Ontolomics-P* also supports users to paste these POIs as described in the "2. Import Data" part.

2. Import Data

This is the first step, in which users could either input a keyword (e.g., a protein name/UniProt ID/noun) or upload a list of POIs. When users prepare their data, they can upload these data from here:

There are two main panels: first, *parameter panel*, users can adjust parameters here; second, *result panel*, many results after users set the parameters will be shown here and users can also download these results.

In the *parameter panel* of "Import Data", there are two choices for users:

a. Type in a keyword. When users choose this option, they can type in a keyword (e.g. TNFSF10) in the "1. Please type in a protein name/UniProt ID/noun" as below.

If users do not input anything, this tool will print "Ontolomics-P detects that you type in nothing!":

b. Upload data. If users choose this parameter, this means users will upload a list of POIs (not just a keyword). There are three options for users:

A. Upload, which means users could upload a a .csv, .txt, .xlsx or .xls file (please refer to "1. Data Preparation" part). Users should select the right format ("1. File format") based on their data and then click "Browse" button ("1.1. Import your data") to import the data.

1.2. Header: this means whether the first row is column names. If true, you should choose this parameter.

1.3. First column: this indicates whether the first column is row names. If true, you should choose this parameter.

1.4. Separator: this is the field separator character. Values on each line of the file are separated by this character.

B. Paste, which means users could paste the POIs in the "1. Paste your data here" area:

C. Load example data, which means users could check the example data if they do not know anything about how to prepare the data.

In the *results panel* of "Import Data", if users don't upload their data, here will show "Ontolomics-P detects that you did not upload your data. Please upload the data, or load the example data to check first." to warn users.

Additionally, users should also recognize which type their data belongs to and choose the right parameter by adjusting the "2. Data type".

3. Ontology-driven Results
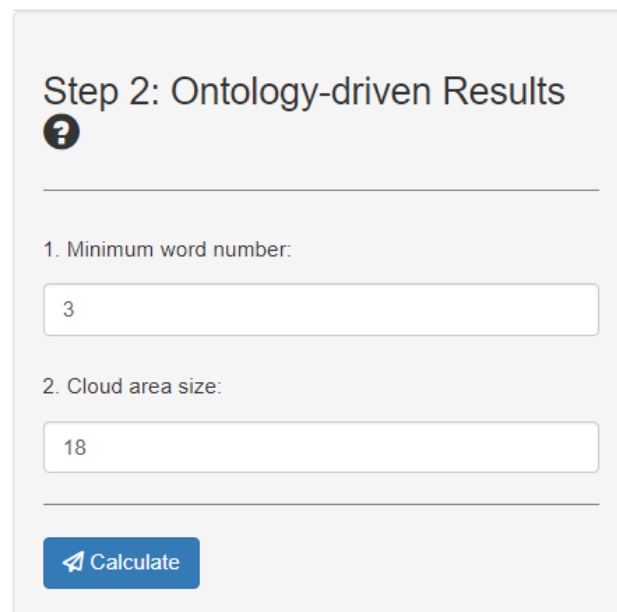
In this step, *Ontolomics-P* built a topics database based on the GO database (please refer to the "Database View" part), aiding in the retrieval of pertinent protein/gene functional information.

3.1 Parameters

If users type in a keyword, the parameter panel will be shown as below:

1. Minimum word number: Those words below this number will not be shown in the cloud plot.

2. Cloud area size: The area size of cloud plot.

If users upload a list of POIs, the parameter panel will be shown as below:

Step 2: Ontology-driven Results ❓

1. Adjusted P value:

> 1

2. GO ID number based on GeneRatio:

> 100

3. Object number for barplot:

> 20

✈ Calculate

1. Adjusted P value: The threshold of BH-adjusted P value for classical GO enrichment results.

2. GO ID number based on GeneRatio: The number of GO IDs with top GeneRatio value from enrichment analysis results, for example, 100 here means this tool uses the GO IDs with the top 100 GeneRatio values to process next simplifyGO function.

3. Object number for barplot: The number of objects (e.g. GO terms or topics) are shown in the barplot, for example, 20 here means the top 20 objects based on the GeneRatio values are shown in the barplot.

3.2 Results of the ontology-driven analysis

After setting these parameters well, then click "calculate" button, the results will appear on the right panel. There are two kinds of results here:

If users type in a keyword (e.g. TNFSF10), here will show: A. Original GO functions, which means this tool retrieves the GO functions of TNFSF10 and plots word clouds of BPs, MFs and CCs respectively. B. Topics analysis results, which means this tool processes topics analysis based on the pre-established topics database.

In the "A. Original GO functions" part:

A.1. Original GO function plots based on the keyword that users type in: Here will show the distribution of the number, word clouds of BPs, MFs and CCs for TNFSF10 respectively.

A.2. Original GO function table based on the keyword that users type in: Here will show the GO functions of TNFSF10 derived from GO database.

In the "B. Topics analysis results" part:

B.1. The correlation plots between the original GOs and those in each topic: This will show the distribution of the semantic similarity between the keyword's (e.g. TNFSF10) GO functions and each topic derived from BPs, MFs and CCs, respectively. Then it also exhibits the histogram of the 125 semantic similarity values.

B.2. The correlation table between the original GOs and those in each topic: This will show the result table of the semantic similarity between the keyword's (e.g. TNFSF10) GO functions and each topic derived from BPs, MFs and CCs, respectively. The semantic similarity values are used to make the figures above (The "B.1. The correlation plots between the original GOs and those in each topic" part).
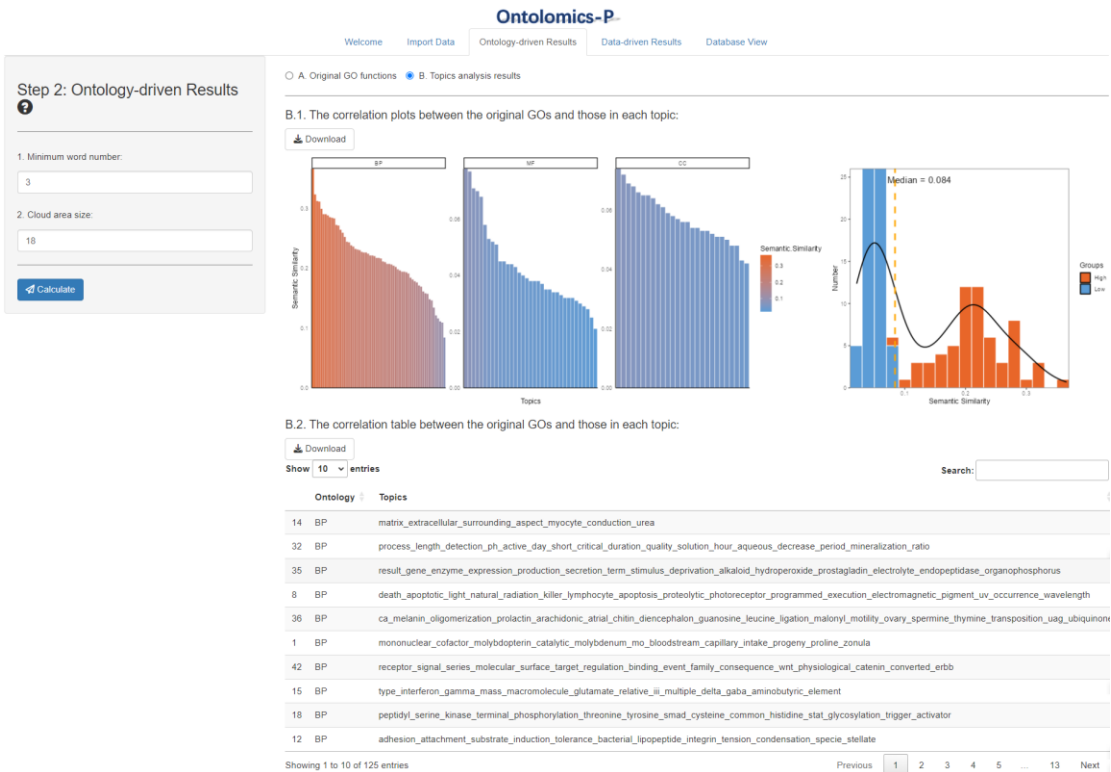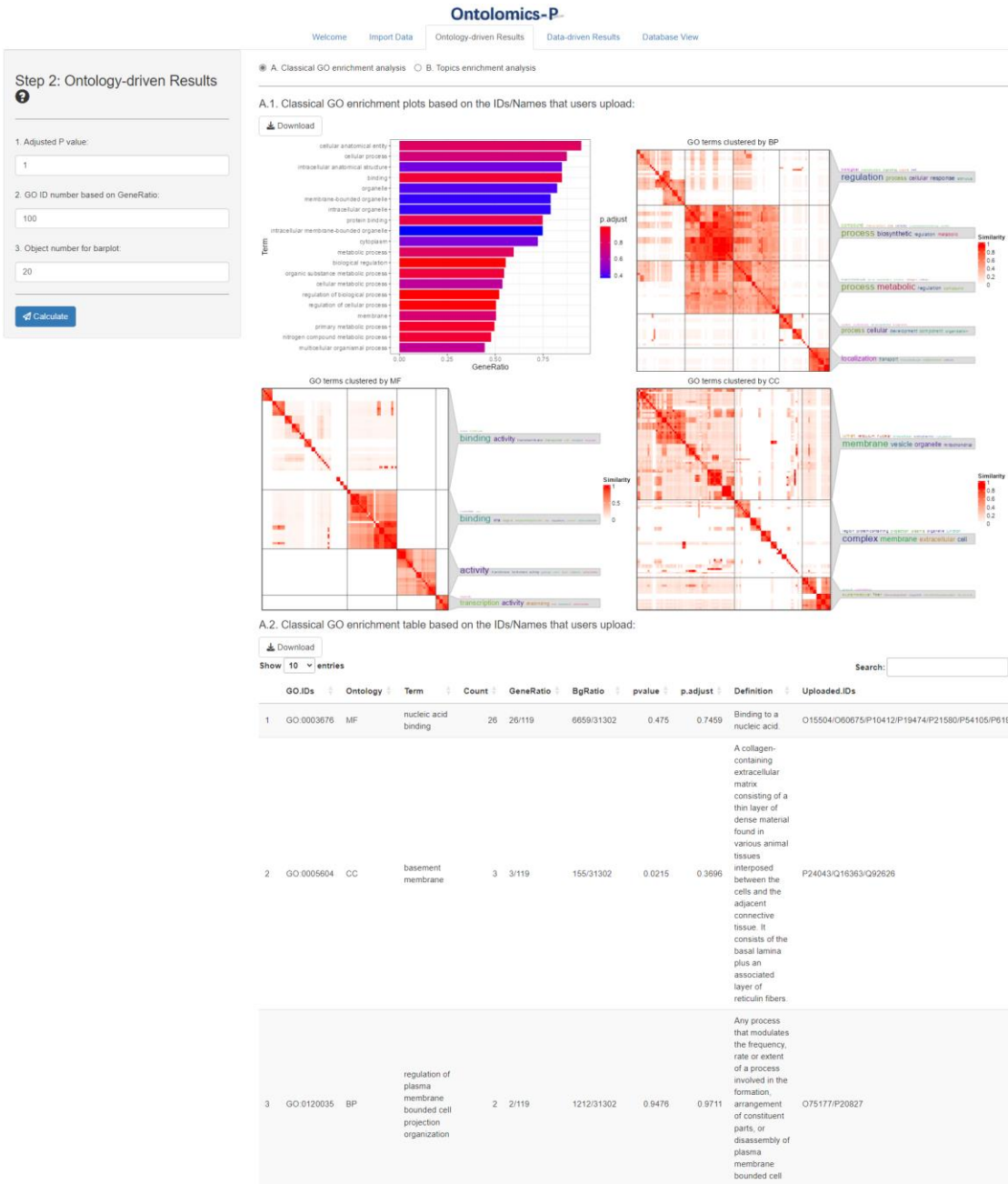
If users upload a list of POIs, here will show: A. Classical GO enrichment analysis, which means this tool retains classical GO enrichment analysis and semantic similarity analysis to simply the results. B. Topics enrichment analysis, which means this tool processes topics enrichment analysis using Fisher's exact test method [11].

In the "A. Classical GO enrichment analysis" part:

A.1. Classical GO enrichment plots based on the IDs/Names that users upload: Here will show the boxplot of the top 20 classical GO enrichment terms and the simplified results of BPs, MFs and CCs of uploaded POIs using heatmaps with word clouds respectively.

A.2. Classical GO enrichment table based on the IDs/Names that users upload: Here will show the classical GO enrichment result table.

In the "B. Topics enrichment analysis" part:

B.1. The topics enrichment analysis plots: Here will show the meteoric plots based on the topics enrichment analysis results of BPs, MFs and CCs of uploaded POIs respectively. Users could adjust the "3. Object number for barplot" parameter to control the topics number of the plot.

B.2. The topics enrichment analysis table: Here will show the topics enrichment analysis result table of BPs, MFs and CCs of uploaded POIs. The meteoric plots above are made based on this table.
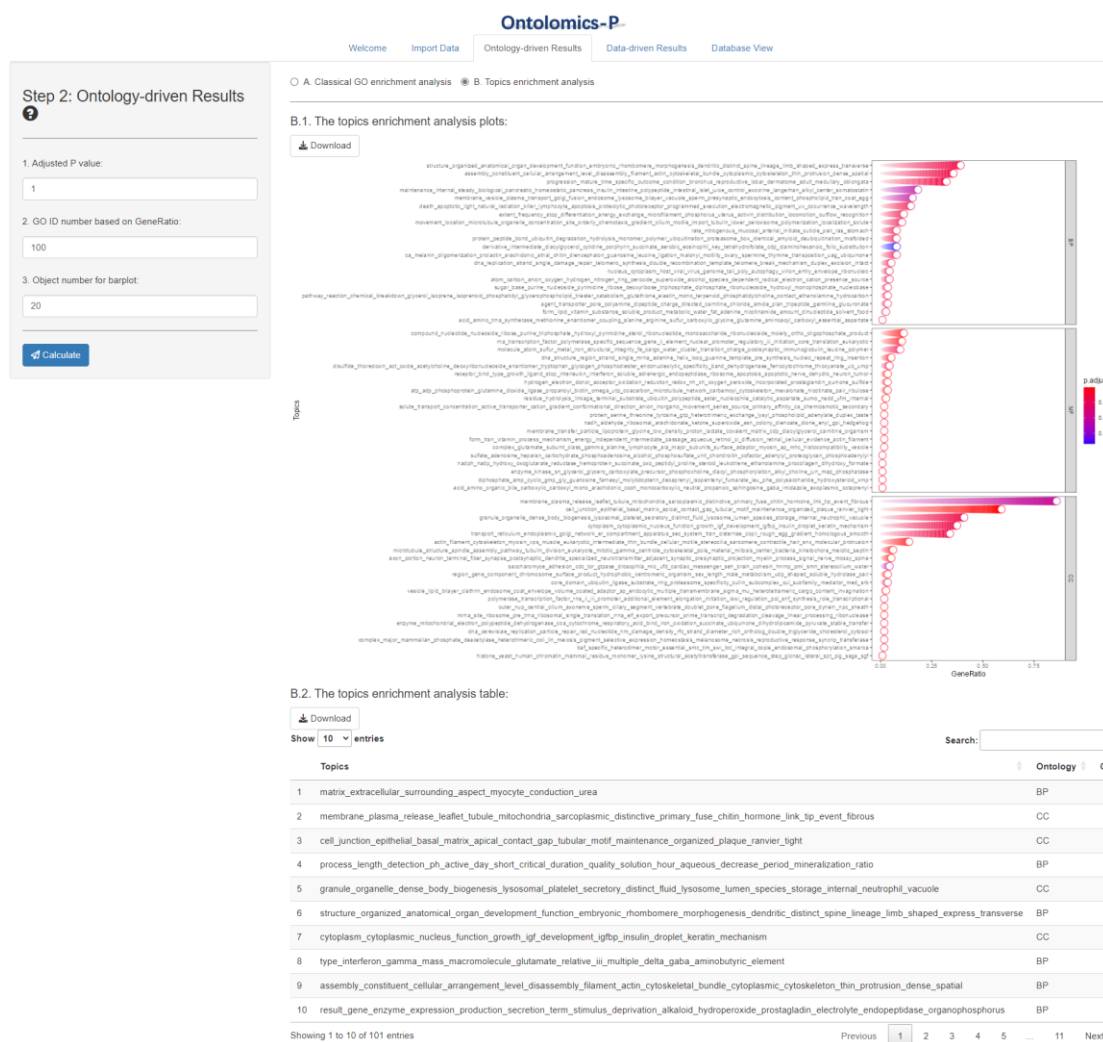
4. Data-driven Results

Here, *Ontolomics-P* seamlessly integrates quantitative proteomic data from ten cancer types stored in the CPTAC database. These expression data were downloaded from https://proteomic.datacommons.cancer.gov/pdc/explore-quantitation-data with PDC study identifiers COAD (PDC000116), BRCA (PDC000120), UCEC (PDC000125), ccRCC (PDC000127), LUAD (PDC000153), HECA (PDC000198), HNSCC (PDC000221), PADA (PDC000270), LSCC (PDC000219) and OV (PDC000360). This integration facilitates the efficient review of the expression profiles of POIs for researchers.

4.1 Parameters

1. Log or not? If true, the expression data will be transformed into logarithms with base 2.

2. Plot type: which kind of plot (Boxplot or Heatmap) users want to show their data. By default, if the protein number is above 10, this tool will only show a heatmap for users to exhibit the data.

3. Z-score or not? If true, the expression data will be normalized by row.

4. Color for volcano plot: Three color names separated with semicolons for the volcano plot.

5. Study names: Ten cancer names integrated in this tool. Users can choose one or more to review the POIs expression.

After selecting suitable methods, users need to click 'Calculate' button, and a popup window will be jumped out to show the selected methods, then click 'OK' button and continue:

4.2 Results of the data-driven analysis

If users type in a keyword (e.g. TNFSF10), In the "A. Figure results" part, here will show relative boxplots and volcano plots with the keyword:
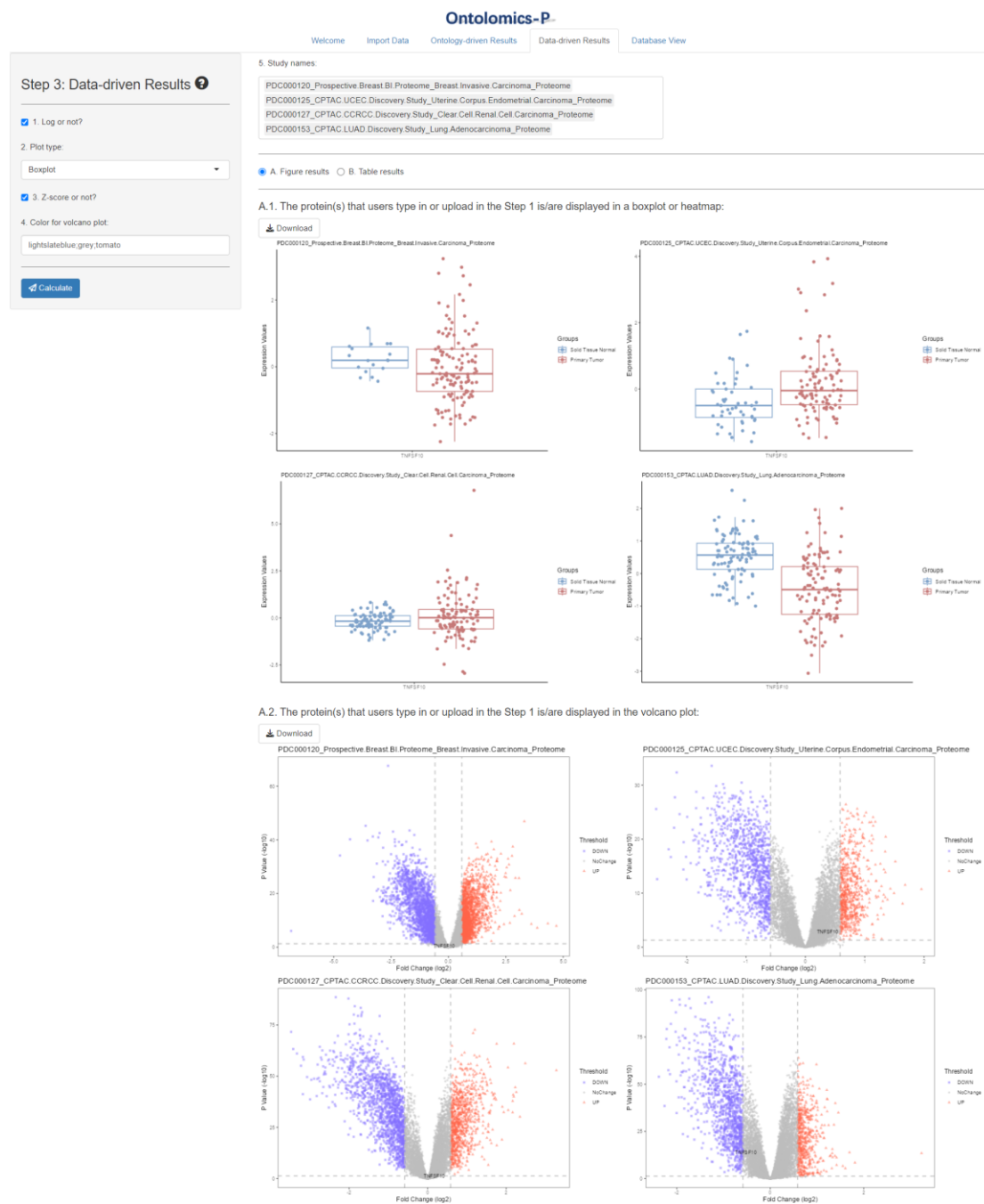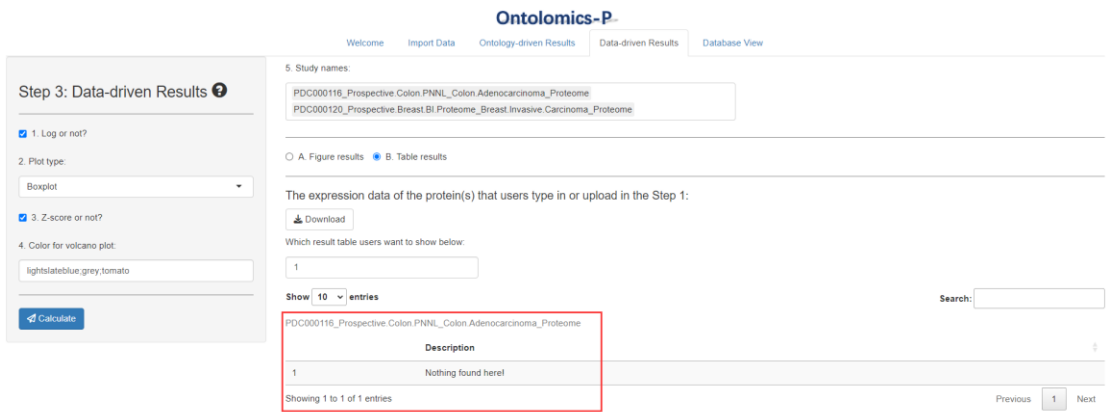
If the protein is not found in the cancer samples, it will show "No protein found" as below:
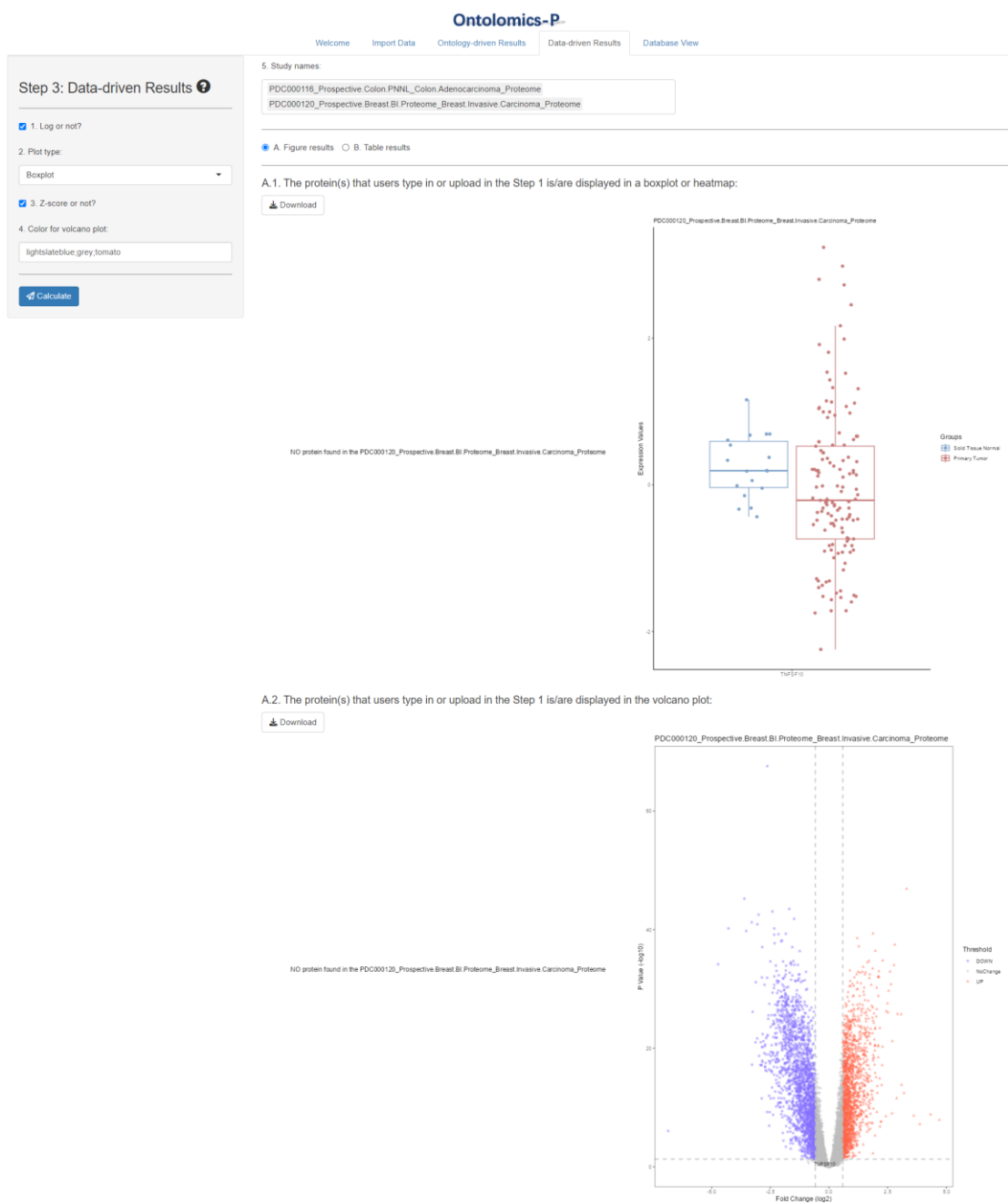
In the "B. Table results" part, it will show the expression table and users can change the "Which result table users want to show below" parameter to check relevant results.
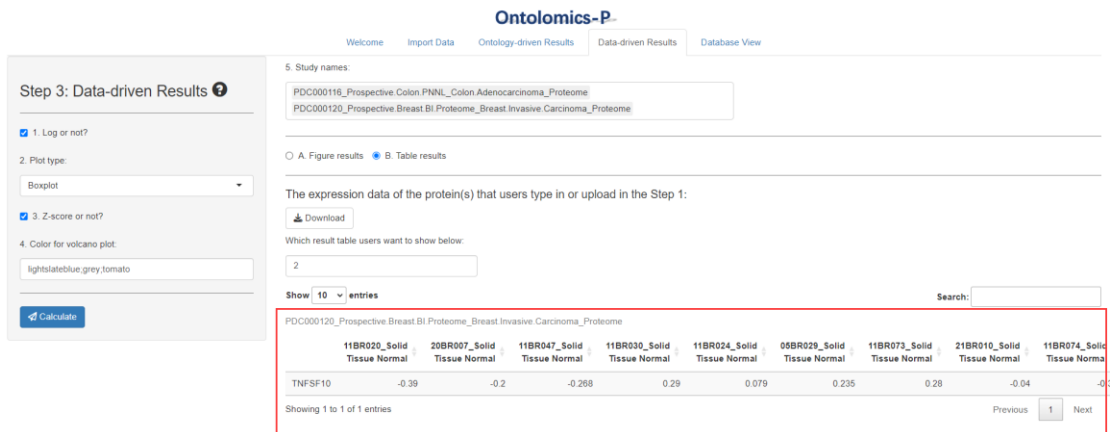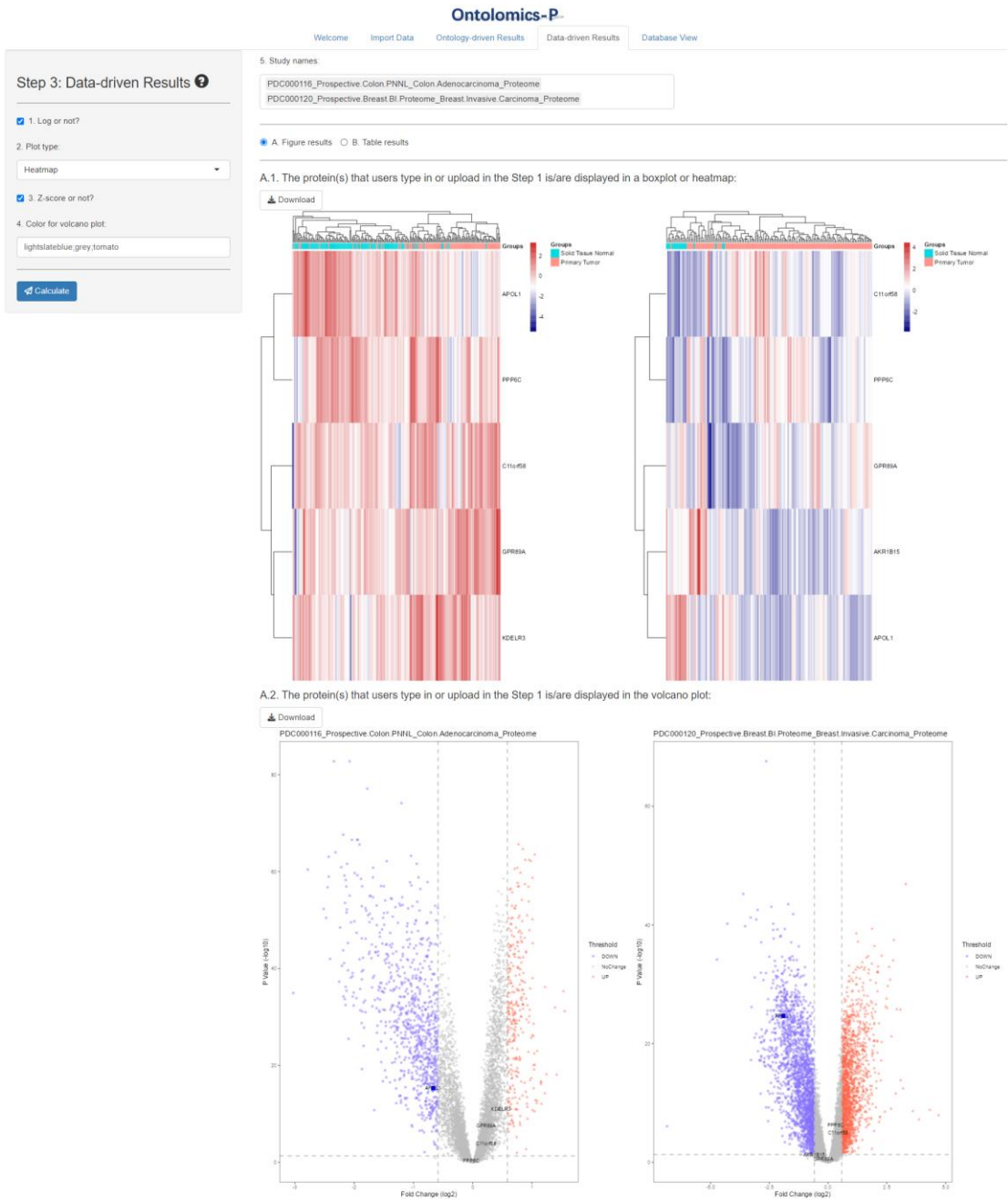

If the protein is not found in the cancer samples, it will show "Nothing found here!" as below:
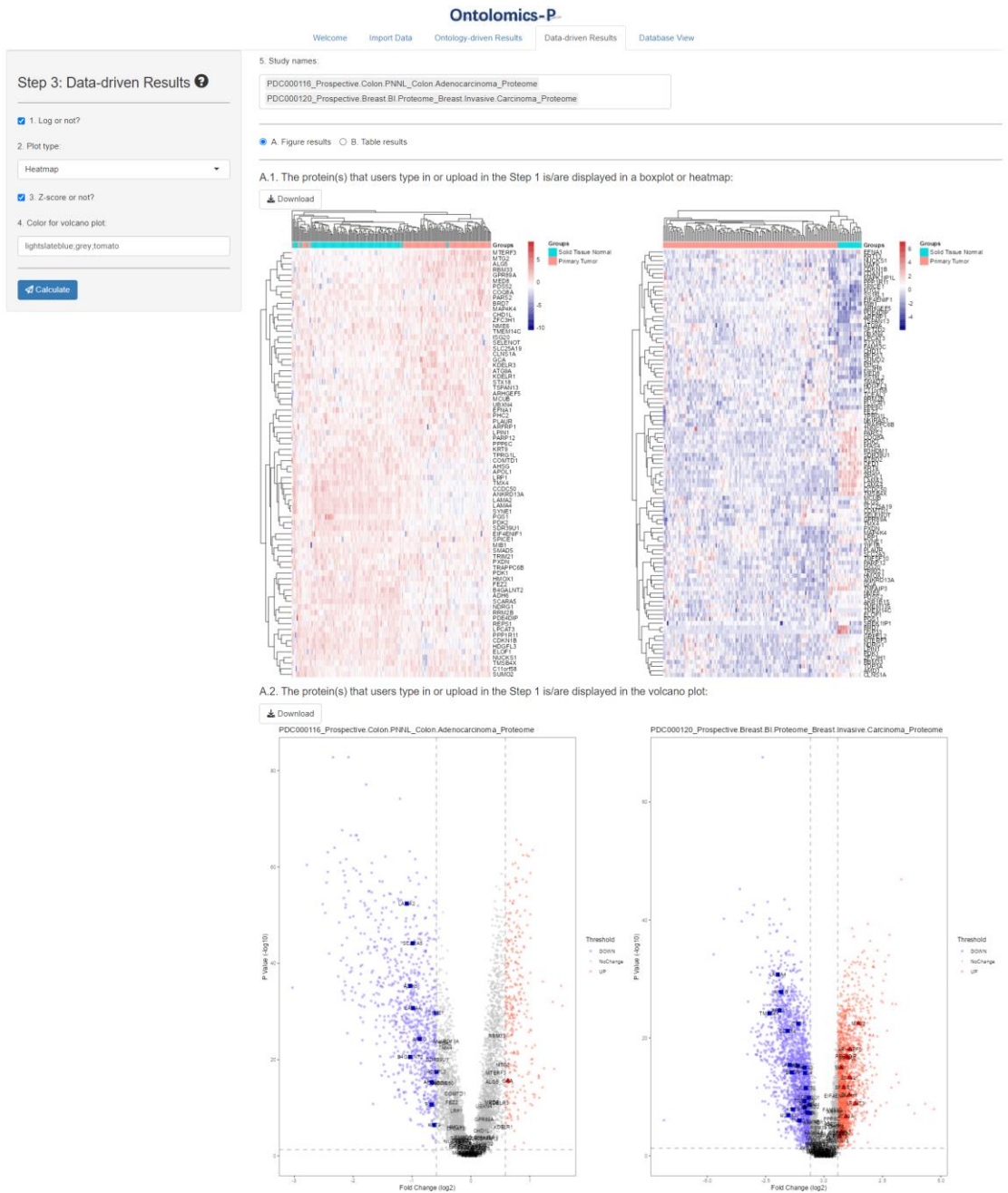
If users upload many but less 10 proteins, In the "A. Figure results" part, here will show relative boxplots and volcano plots:
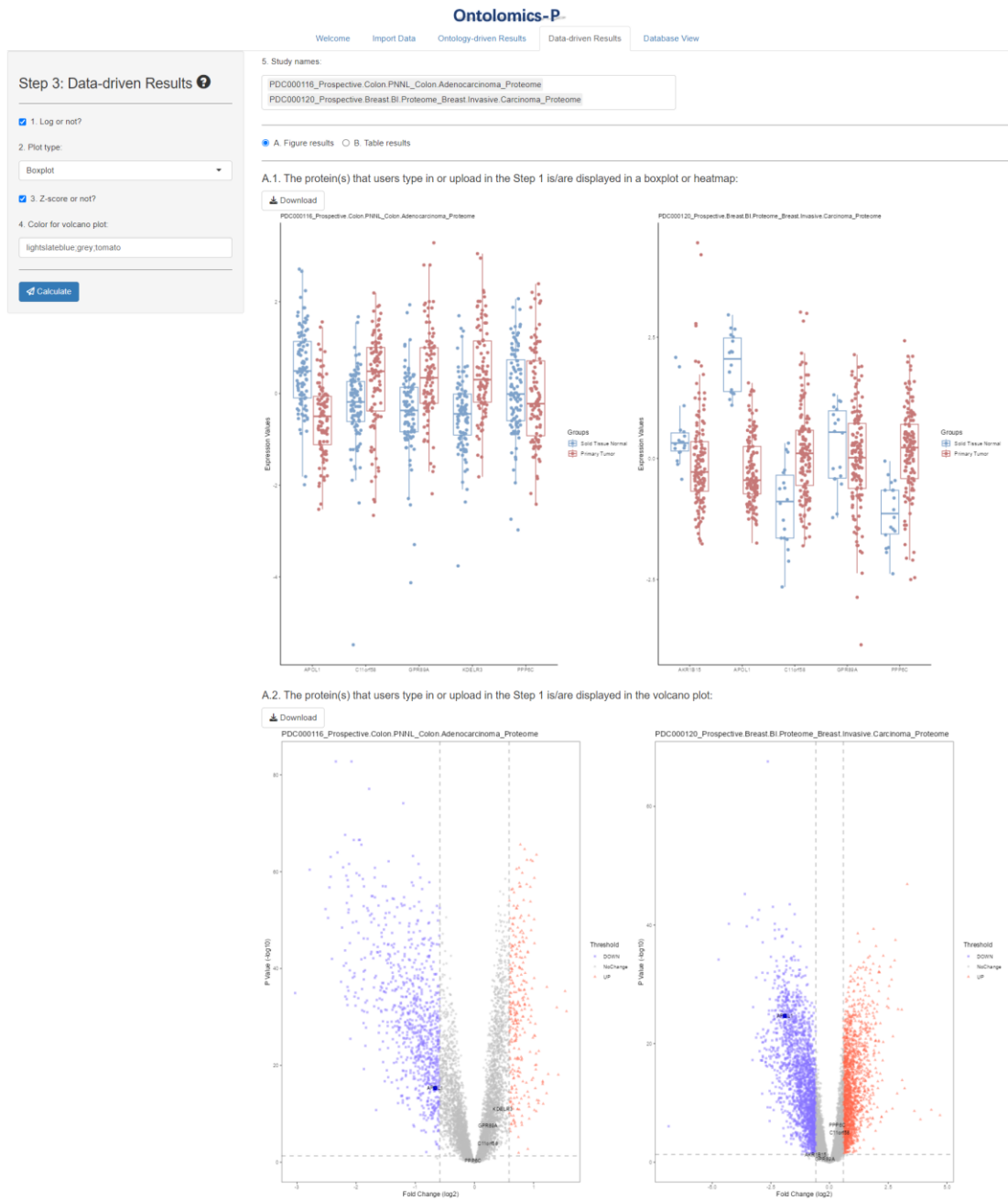
Or users can change the "2. Plot type" parameter to the heatmap type:

If users upload a list of proteins (>10), In the "A. Figure results" part, here will show relative heatmaps and volcano plots:

5. Database View

This step will show the databases integrated in *Ontolomics-P*.

1. Topics analysis database: This tool built a topics database based on the GO database using the Latent Dirichlet allocation (LDA) model [12] and Users can check the database from here.

2. GO terms used for topics analysis: This database was implemented using GO.db package (Version 3.17.0, https://doi.org/doi:10.18129/B9.bioc.GO.db).

6. How to run this tool locally?

Ontolomics-P is an open source software for non-commercial use and all codes can be obtained on our GitHub: https://github.com/wangshisheng/Ontolomics-P. If users want to run *Ontolomics-P* on their own computer, they should operate as below:

As this tool was developed with R, you may:

a) Install R. You can download R from here: https://www.r-project.org/.

b) Install RStudio. (Recommendatory but not necessary). You can download RStudio from here: https://www.rstudio.com/.

c) Check packages. After installing R and RStudio, you should check whether you have installed these packages (shiny, shinyjs, shinyWidgets, shinyBS, DT, data.table, ggsci, ggplot2, ggrepel, patchwork, dplyr, GOSemSim, openxlsx, clusterProfiler, simplifyEnrichment, tidytext, ggwordcloud, cowplot, grid, ggpubr, impute, pheatmap, ggplotify, limma, qvalue). You may run the codes below to check them:

```
if(!require(pacman)) install.packages("pacman")
pacman::p_load(shiny, shinyjs, shinyWidgets, shinyBS, DT, data.table, ggsci, ggplot2,
ggrepel, patchwork, dplyr, GOSemSim, openxlsx, clusterProfiler, simplifyEnrichment, tidytext,
ggwordcloud, cowplot, grid, ggpubr, impute, pheatmap, ggplotify, limma, qvalue)
```

Please note, if you find some packages cannot be installed directly using the above command, you can find them in the GitHub source and install them by, for example:

```
library(devtools)
install_github("wangshisheng/Ontolomics-P")
```
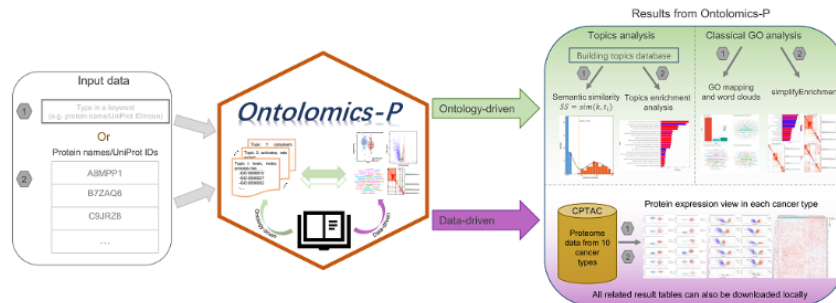
d) Run this tool locally

```
if(!require(Ontolomics-P)) devtools::install_github("wangshisheng/Ontolomics-P")
library(Ontolomics-P)
Ontolomics-P_app()
```

Then Ontolomics-P will be started as below (same as the online version), and the detailed operation about Ontolomics-P can be found in the Supplementary Notes part 1-5 above:

Ontolomics-P

Welcome | Import Data | Ontology-driven Results | Data-driven Results | Database View

~~ Dear Users, Welcome to Ontolomics-P ~~

Ontolomics-P is a web-based tool, which possesses the core functions, including:

1. Topic analysis integration. Leveraging natural language processing, this feature allows researchers to identify and categorize various topics within a collection of documents. It builds a topics database based on the Gene Ontology (GO) database, aiding in the retrieval of pertinent protein/gene functional information;

2. Topics enrichment analysis implementation. Utilizing Fisher's exact test method, Ontolomics-P conducts topics enrichment analysis to identify significant associations and enrichments within the dataset;

3. Inclusion of classical GO enrichment analysis and semantic similarity analysis. To enhance practicality, Ontolomics-P retains classical GO enrichment analysis and semantic similarity analysis, offering diverse analytical approaches;

4. Integration of quantitative proteomic data from diverse cancer types. Ontolomics-P seamlessly integrates quantitative proteomic data from ten cancer types stored in the CPTAC database. These include colon adenocarcinoma (COAD), breast cancer (BRCA), uterine corpus endometrial carcinoma (UCEC), clear cell renal cell carcinoma (ccRCC), lung adenocarcinoma (LUAD), hepatocellular carcinoma (HECA), head and neck squamous cell carcinoma (HNSCC), pancreatic ductal adenocarcinoma (PADA), lung squamous cell carcinoma (LSCC), and ovarian cancer (OV).

In addition, this tool supports both online access and local installation. The source codes and installation instructions can be available in the GitHub repository: https://github.com/wangshisheng/Ontolomics-P under an MIT license.

Finally, Ontolomics-P is developed by R shiny (Version 1.6.0), and is free and open to all users with no login requirement. It can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. We would highly appreciate that if you could send your feedback about any bug or feature request to Shisheng Wang at shishengwang@wchscu.cn.

Friendly suggestions:

a) Open Ontolomics-P with Chrome, Mozilla Firefox, Safari or Firefox;

b) The minimum operating system specifications are: RAM 4GB, Hard drive 100 GB;

c) The monitor resolution (>= 1920x1080) is better.

*^_^ Enjoy yourself in Ontolomics-P ^_^*

III. References

[1] Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell. 2019;177:1035-49 e19.

[2] Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, Anurag M, et al. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell. 2020;183:1436-56 e31.

[3] Dou Y, Kawaler EA, Cui Zhou D, Gritsenko MA, Huang C, Blumenberg L, et al. Proteogenomic Characterization of Endometrial Carcinoma. Cell. 2020;180:729-48 e26.

[4] Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell. 2019;179:964-83 e31.

[5] Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, et al. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell. 2020;182:200-25 e35.

[6] Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, et al. Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell. 2019;179:561-77 e22.

[7] Huang C, Chen L, Savage SR, Eguez RV, Dou Y, Li Y, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell. 2021;39:361-79 e16.

[8] Cao L, Huang C, Cui Zhou D, Hu Y, Lih TM, Savage SR, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell. 2021;184:5031-52 e26.

[9] Satpathy S, Krug K, Jean Beltran PM, Savage SR, Petralia F, Kumar-Sinha C, et al. A proteogenomic portrait of lung squamous cell carcinoma. Cell. 2021;184:4348-71 e40.

[10] Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell. 2016;166:755-65.

[11] Jung SH. Stratified Fisher's exact test and its sample size calculation. Biom J. 2014;56:129-40.

[12] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3:993-1022.