EnrichMiner: a biologist-oriented web server for mining biological insights from functional enrichment analysis

Xian Liu[1,#,*], Kaikun Xu[2,#], Xin Tao[2,3], Xiaochen Bo[1,*], Cheng Chang[2,4,*]

1 Institute of Health Service and Transfusion Medicine, Beijing 100850, China

2 State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China

3 School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

4 Research Unit of Proteomics Driven Cancer Precision Medicine, Chinese Academy of Medical Sciences, Beijing 102206, China

#These authors contributed equally: Xian Liu and Kaikun Xu

* To whom correspondence should be addressed:

Xian Liu, Email: liux.bio@gmail.com

Xiaochen Bo, Email: boxc@bmi.ac.cn

Cheng Chang, Email: changchengbio@163.com

Abstract: Functional enrichment analysis has been widely used to help researchers obtain biological insights from -omics data. However, the results are often redundant and difficult to digest. The key is developing tools to help users explore the relationships between the enriched terms, remove the redundant terms, and finally select representative terms. However, existing tools hardly make a good integration between enrichment analysis and representative terms selection in a biological-friendly manner. Here, we developed a biologist-oriented web server named EnrichMiner to provide a one-stop solution. It is a complete analysis pipeline from a gene list or a ranked gene table to published-style figures. More importantly, it provides user-friendly interfaces and rich interactive operations to help users explore the term relationships and remove redundancy. EnrichMiner has been integrated into the ExpressVis platform, and is freely accessible at https://omicsmining.ncpsb.org.cn/ExpressVis/EnrichMiner and does not require login.

Keywords: functional enrichment analysis, interactive visualization, redundancy reduction, directed acyclic graph, multi-omics

1 Introduction

Although omics technologies, such as RNA-seq and mass spectrometry, have become widely used in biomedical research, mining biological insights from these omics data remains a huge challenge. One of the routine analysis methods is functional enrichment analysis. At present , the most commonly used enrichment analysis methods are Over Representation Analysis (ORA)[1] and Gene Set Enrichment Analysis (GSEA)[2]. ORA takes a gene/protein list of interest as input and iteratively tests whether one gene set (such as all the genes in a pathway) contains disproportionate genes of the input. Differently, GSEA takes a gene table with ranked metrics as input and tests whether genes of a gene set accumulate at the top or bottom of the full gene vector ordered by quantitative values[3]. Many excellent tools have been developed for the two enrichment methods, such as DAVID[4], g:Profiler[5], Enrichr[6] and WebGestalt[7]. Meanwhile, similar analysis methods such as Metabolite Set Enrichment Analysis (MSEA)[8]and Kinase–Substrate Enrichment Analysis (KSEA)[9] have also been developed to identify the significant metabolites or kinase's activity.

However, digesting the enriched terms remains a big challenge due to the high redundancy. The background gene set library usually contains large number of gene sets with overlapping genes and the ORA and GSEA methods consider each term independently. Thus, they usually output the enriched terms with high redundancy. Some tools have been developed to remove the redundancy of enriched terms to make them more comprehensive and digestible. There are some R or Python packages, such as topGO[10], GOSemsim[11] and GOMCL[12], aiming for removing the redundancy of enriched terms by measuring similarities between GO terms. But they are not easy to use for biologists with few programming skills and their results are not user-friendly. There are also some web-based tools for this goal. For example, REVIGO[13] enables to pre-filter the enriched terms using a cluster algorithm of semantic similarity to find a representative subset and visualize the clustering results with a network map. But it did not include the ORA or GSEA function. Another popular tool, Metascape[14], pre-clustered the enriched terms by Kappa score, and selected the term with the smallest P-value in each cluster as the representative one. However, Metascape only supports ORA analysis and the figures provided by Metascape are non-interactive static figures. Users cannot interactively explore the enrichment results. And it is not easy for users to explore the similarity between terms due to the lack of visualization of overlapping composition between terms within a cluster. Thus, there are few suitable tools for a comprehensive analysis from enrichment analysis to representative terms selection.

Here, we propose a biologist-oriented web server, named EnrichMiner, to solve the above limitations. It implements two ways with rich interactive operations to visualize the relationships between the terms, reduce the redundancy and select the representatives. One is constructing a network according to pairwise term similarity, clustering the terms in the network with user-selected community detection method. The other is visualizing the hierarchical structure of the enriched terms through directed acyclic graph (DAG). Additionally, both similarity network map and DAG provides interactive operations to overview the relationships between all the terms, then dive into the relationships between sub terms, and finally view the detail of one term. Along with the maps, there are detailed

tutorial videos to help users. And EnrichMiner provides interactive operations with the modules in our previous platform ExpressVis[15]. Most importantly, EnrichMiner contains a complete pipeline for extracting biological insights from enrichment analysis. It supports both ORA and GSEA. After uploading a gene list or a gene table with rank metrics (such as the differential analysis results), users can perform ORA or GSEA, explore the relationships of the enriched terms, select the representatives, and finally visualize them in 7 kinds of downloadable figures. We present one re-analyzed case study to demonstrate the features and capabilities of EnrichMiner.

2 Results

2.1 Overview of EnrichMiner

EnrichMiner is a complete pipeline for mining biological insight from omics data. As shown in Fig. 1, its workflow can be divided into three steps: (1) statistical calculation, (2) selection of representative terms after exploring term relationships and removing redundancy, (3) visualization of the representative terms.

In the statistical calculation step, EnrichMiner provides multiple options for users. At the method level, it supports ORA analysis and GSEA analysis. ORA takes a gene list as input and GSEA takes a table with ranked metrics as input. At gene set library level, it supports rich gene set libraries that are stored in the database and those uploaded by the users (see Methods for details). At data level, it supports transcriptomics and proteomics data.

Selection of representative terms is the key of EnrichMiner. It aims to liberate researchers from the complicated comparison and empirical selection of enriched terms by supporting two strategies for selecting representative terms: (1) Construct a network based on the pairwise similarity between terms and clustering terms with community detection algorithms; (2) Manually select terms of interest by terms relationship visualization through a DAG. More importantly, rich interactive visualization operations are available to explore the ORA and GSEA results.

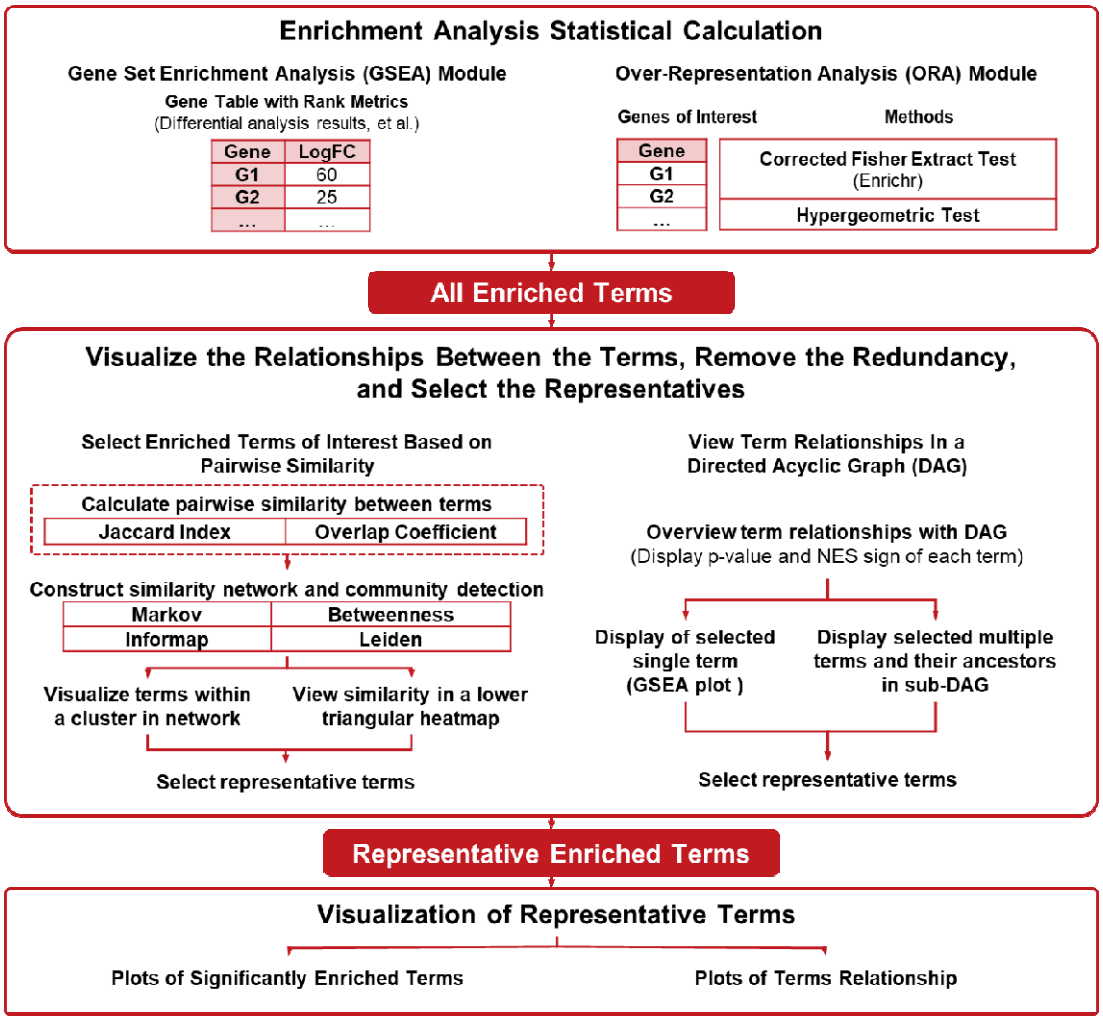

Figure 1. The overview of EnrichMiner workflow. Users can upload a gene table with ranked metrics (for GSEA) or a gene list (for ORA) to calculate the statistical significance of all gene set terms. The most representative terms can be selected with two redundancy reduction methods: community detection of similarity network and term relationship exploration using DAG. Rich interactive operations are available to explore representative terms and relationships among them.

2.2 Reduce redundancy and select representative terms

Since in the commonly used gene set libraries like Gene Ontology (GO), a large number of terms contains overlapping genes. The high redundancy of the terms hampers biological interpretations of the enrichment results. In EnrichMiner, to help users explore the terms' relationships and select representative gene set terms, we implement two methods: one is based on terms pairwise similarity and the other is based on parent-child relationships with DAG (Fig. 2).

2.2.1 Reduce redundancy and select representative terms based on terms pairwise similarity.

Gene sets with similar gene composition are considered to represent similar biological

functions. As shown in Fig. 2A and Fig. S1A, EnrichMiner constructs a similarity network of the enriched terms, where nodes represent terms and edges represent pairwise-term similarity. Then, the enriched terms are clustered into multiple groups through community detection algorithm (Markov clustering, Informap method or Leiden method). Users can interactively select similarity definition (Jaccard index or overlap similarity) and cutoff to construct the network and choose community detection algorithm to cluster the terms. EnrichMiner provides interfaces to visualize clusters and terms with a cluster. The clustering results are visualized in a network map, a triangular matrix map and summarized table. In the network map, terms with the smallest p-value within each group are selected as represented term and they are displayed as group nodes. Term nodes are displayed with different colors according to the groups they belong to. Size of the node represents gene number in the term and width of the edge represents similarity (Fig. S1B). Although the network map made by EnrichMiner is downloadable for publishing with little modification, users can utilize the following operations to modify the network map: 1) Click to refresh the layout of the network; 2) Move the terms in a group simultaneously; 3) Change or move group label; 4) Change group color (the nodes color changes accordingly); 5) Change the attributes mapped to node size and edge width. In the triangular matrix map, the color represents similarity of the term pairs. The summarized table lists clusters information, such as number of terms, the largest term and the term with smallest p-value within each cluster.

The network map, triangular map and summarized table provides operations for users to navigate to the interfaces for exploring the terms within a cluster. EnrichMiner provides three interfaces: a table, a term similarity network and term-gene network. The table list terms enrichment information, such as p-value, genes in the term. The terms similarity network map displays terms similarity, in which node represents term, edge represents similarity. The term-gene network map displays terms and genes, in which one type of node represents term, another type of node represents genes, the edge between gene node and term node represent that the gene is in the term.

2.2.2 Reduce redundancy and select representative terms based on terms parent-child relationships.

Gene sets in some gene set libraries, such as gene ontology and Reactome pathways, have parent-child relationships and they can be represented as DAG. The lower-level terms in DAG are more specific while high-level ones are often general. If nodes in a path are mostly significantly enriched, the corresponding biological function are more robust and notable. After obtaining the enriched terms, EnrichMiner constructs the DAG of them and their ancestors. For the number of enriched terms and their ancestors may be very large and so is the DAG map, visualizing DAG appropriately in web is a challenge. EnrichMiner provides three options: one visualizes the whole DAG, one visualizes the sub-DAG of the enriched terms and the nodes in the shortest path between them and the root, one only visualizes the sub-DAG of the enriched terms. In the DAG map, node represents term and directed edge represents the two edges have parent-child relationship. EnrichMiner provides rich operations to explore the DAG map: users can drag, move whose DAG map

or one DAG node, search and locate terms of their interest. To make DAG map more interactive, we designed carefully the node box. The box is divided into five parts (Fig. S1C): (1) the outer border. (2) the top rectangle box, (3) the middle rectangle box, (4) the lower left check-box, (5) the lower right check-box. The color of outer border indicates NES of GSEA results (NES < 0, green; NES > 0, brown). The top rectangle box displays the term ID and users can click it to navigate to the gene ontology website to see the detail of the term. The middle rectangle displays the term name, and its background color indicates the enrichment significance (i.e., the darker red indicates the smaller P-values). In GSEA module, when users click on the term name, the corresponding GSEA plot and gene composition list will be displayed. When users click lower left box, a sub-DAG will be displayed to show all selected terms and their ancestor nodes. Through the above interactive operations, users can get an overall impression of the enriched terms and their relationships, and finally determine which terms can be selected as representative terms. Then users can check lower right box to select it as a representative term and add it into the candidate table.

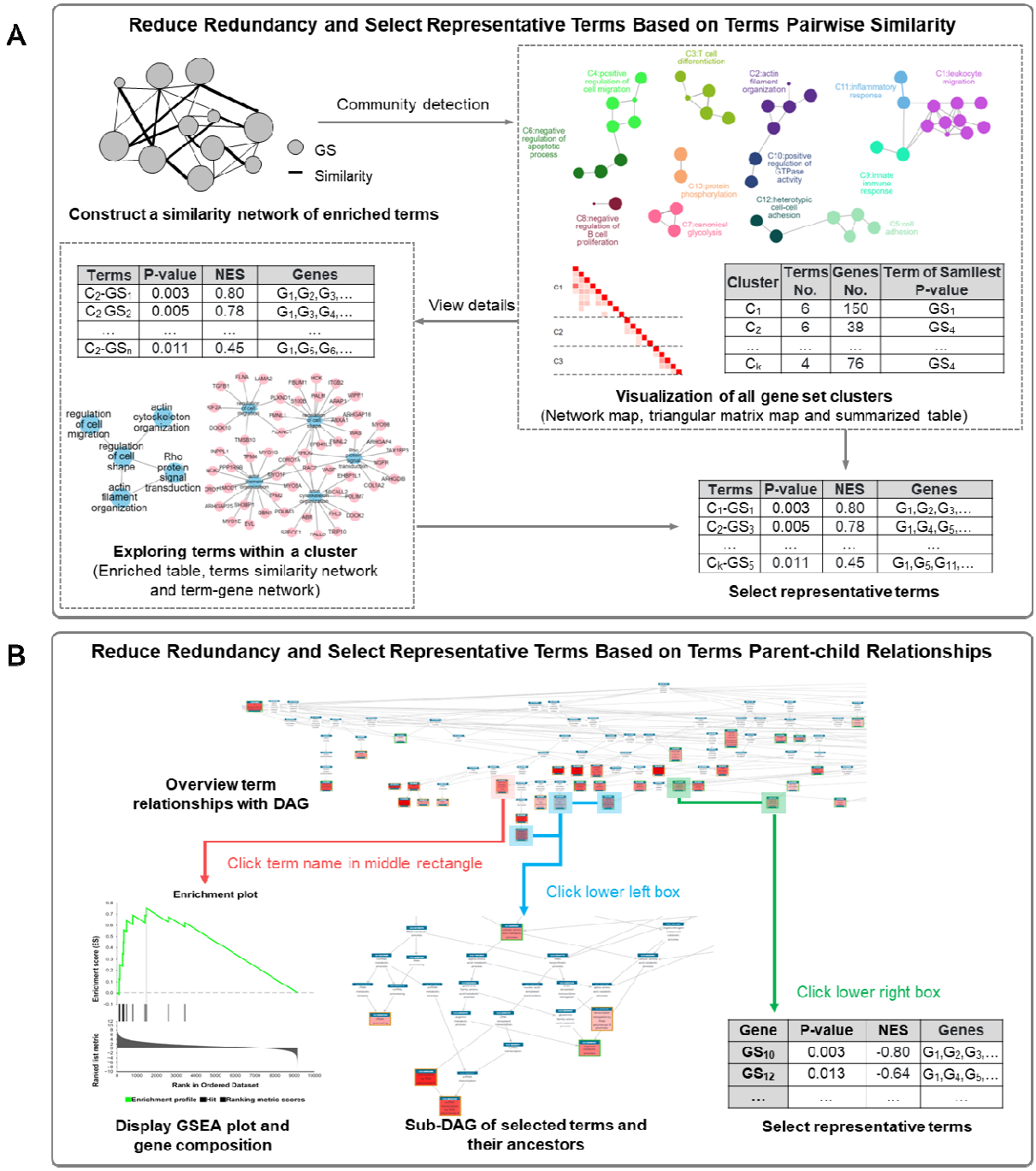

Figure 2. Select representative terms via community detection of pairwise similarity network and DAG exploration. (A) Similarities of the significantly enriched terms are extracted based on pre-calculated similarity data matrix and used to construct similarity network. The network can be divided into clusters using community detection algorithms. The results are displayed in a network map, a triangular matrix map and a summarized table. Terms within a cluster can be shown through an enriched table, a terms similarity network and a term-gene network. (B) For GO analysis, significantly enriched terms are highlighted in a DAG (red indicates the range of P-value, brown/green indicates positive/negative of NES in GSEA module). Users can locate terms of interest easily.

2.3 Visualization of the representative terms

The visualization of the representative terms is composed of two types: those display enriched significances and those display relationship among terms (Fig. 3). In both GSEA and ORA modules, interactive scatter plot, bar plot and lollipop plot are generated to show -Log_{10}(P-value) of selected terms (Fig. 3A). Additionally, in GSEA module, multi-terms GSEA plot and Rank list of GSEA table are generated to show the normalized enriched score (NES). Term relationships are displayed via upset plot and "gene set-to-gene" heatmap (Fig. 3B). In upset plot, the cardinality of single terms as well as term combinations with more than 3 gene sets are presented. The total size of each set is represented on the left bar plot. Every possible intersection is represented on the bottom plot, and the number in each subset is shown on the top plot. In heatmap, genes included in each term will be filled with red color.

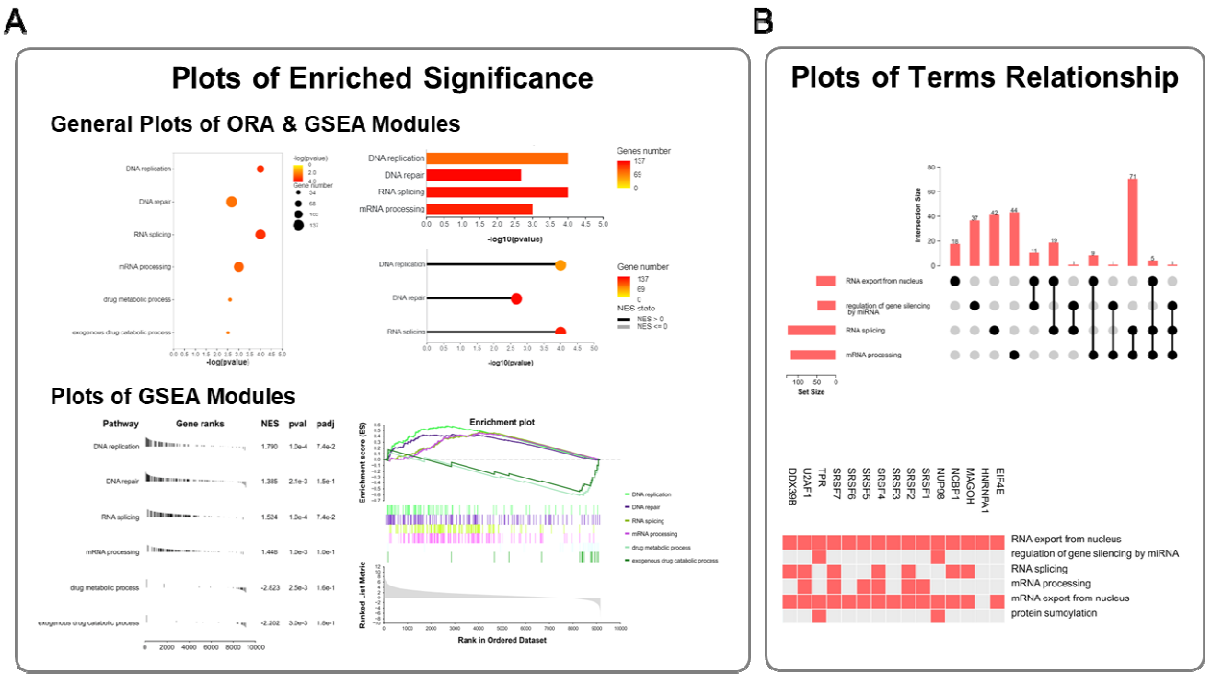

Figure 3. Visualization of EnrichMiner. (A) The statistical significances of enriched terms. (B) The relationships among terms.

2.4 Case study I: A re-analysis of published HCC proteomics dataset

To demonstrate the capabilities of EnrichMiner, we re-analyzed a published HBV-related HCC proteomics dataset[16], which contains data of 159 pairs of tumors (T) and adjacent non-tumor (N). The 882 down-regulated proteins in tumor with tumor thrombus from Supplymentary Table 3 in that paper are used for analysis in this study[16].

As shown in Fig. 4A, when we used KEGG pathway database as the background in the ORA module, terms chosen in the original Fig. 3F (*Biosynthesis of amino acids*, *Glyoxylate*

and dicarboxylate metabolism and *Carbon metabolism*) of Gao et al. paper[16] are clustered in C2. In addition, 18 proteins are overlapped in *Carbon metabolism* and *Glyoxylate and dicarboxylate metabolism*.

When we used logFC(T/N) as ranked metrics in the GSEA module, similar conclusions were obtained (Fig. 4B). Fatty acid metabolism related terms (*Monocarboxylic acid metabolic process*, *Fatty acid metabolic process*, *Fatty acid biosynthetic process* and *Long-chain fatty acid biosynthetic process*) and P450 related terms (*Arachidonic acid metabolic process* and *Omega-hydroxylase P450 pathway*) are significantly enriched in tumor. While, RNA processing related terms (*RNA metabolic process*, *RNA processing*, *mRNA processing*, *mRNA 3'-end processing*) and DNA repair related terms (*Cell response to DNA damage stimulus*, *DNA repair*, *Double-strand break repair*, *Double-strand break repair via homologous recombination* and *Double-strand break repair via break-induced replication*) are significantly enriched in non-tumor.

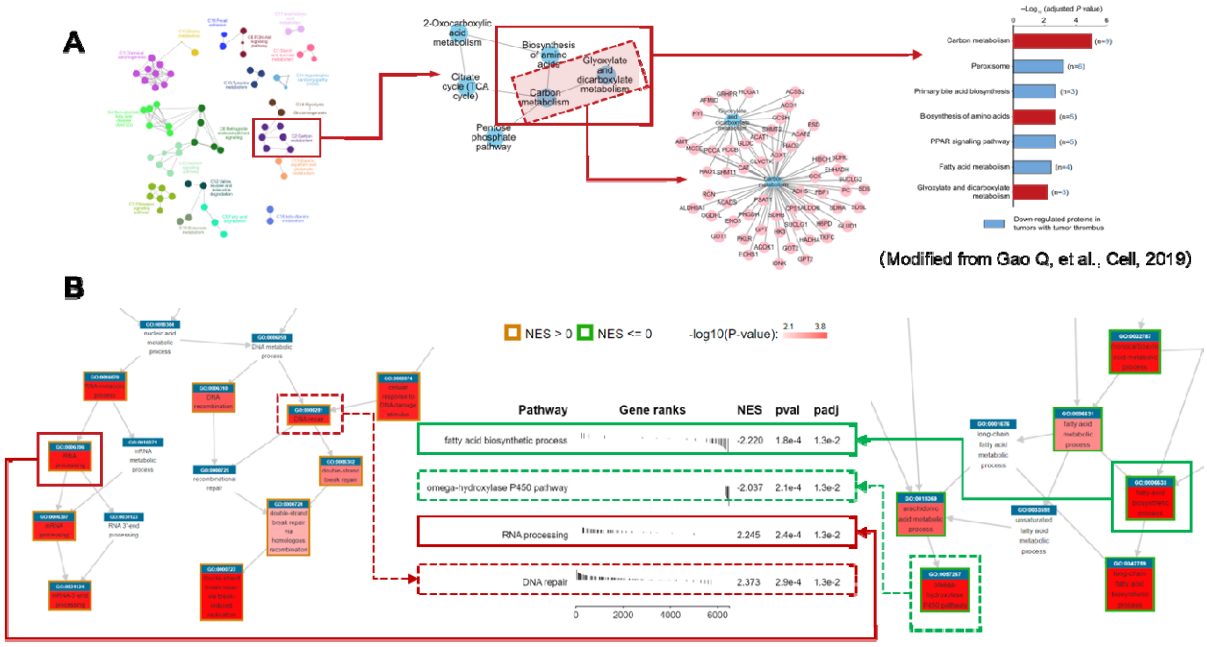

Figure 4. Enrichment results of published HCC Proteomics study (Gao et al.'s cohort). (A) ORA enriched terms of KEGG pathway library based on down-regulated proteins between tumors and adjacent non-tumors. (B) GSEA enriched terms of GO Biological Process gene set database based on logFC between tumors and adjacent non-tumors.

3 Discussion

Functional enrichment analysis has been used to extract biological insights from -omics data for more than two decades. Although researchers have developed many sophisticated statistical methods, extracting biological insights still requires domain experts' experience and intuition. To take advantage of experts' experience and intuition, we should develop tools as a visual analytical platform. EnrichMiner contains well-designed visualization interfaces and rich interactive operations to help users overview the relationships of the

enriched terms, explore the relationships of the sub terms, and dive into the details of one term. With EnrichMiner, users can have a better understanding of their functional enrichment results, select representative terms and generate publishable figures.

In summary, with the well-designed visualization interfaces and rich interactive operations, EnrichMiner is expected to be a useful tool for extracting biological insights from functional enrichment results in an efficient and convenient way. Table 1 lists all the major features of EnrichMiner compared with five other similar web servers. First, EnrichMiner is a complete pipeline, including ORA and GSEA for enrichment analysis, similarity network and DAG for exploring the relationships of the enriched terms and selecting the representatives, and rich interactive and visualization operations for extracting biological insights from representative terms. Second, many user-friendly interactive visualization functions are implemented in EnrichMiner due to the latest web development techniques we used and it can be considered as one of the most important features of EnrichMiner.

However, the current version of EnrichMiner has some limitations: the gene set libraries are not enough for the increasing omics studies and an efficient and automatic data synchronization procedure should be considered. In the future, we plan to continue updating EnrichMiner to overcome these limitations. Meanwhile, as large language models (LLM, such as ChatGPT) are keep improving, the way of human-computer interaction will change, and the development of corresponding software tools may experience a paradigm shift. Microsoft is bringing ChatGPT technology to the Office software (such as Word, Excel and Outlook). In the future, we will also pay attention to the integration between EnrichMiner and LLM for better interaction and better user experience.

4 Methods

4.1 Webserver Architecture

The architecture is similar with that of our previous work ExpressVis. The font-end focuses on visualization and human-computer interaction, the backend focuses on computation. The front-end is implemented with Angular and the server is implemented with Django. EnrichMiner requests Enrichr's results directly from the browser.

4.2 Background Gene Set Library

For ORA and GSEA, gene set libraries play critical roles in enrichment analysis, which are made of a set of related gene lists associated with a functional term such as specific biological processes, pathway membership, or other features.

EnrichMiner includes multiple gene set libraries of model organisms, such as *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. The gene set libraries are three sources: MSigDB, Enrichr and those collected from the gene set databases directly. Additionally, EnrichMiner supports user-defined gene set libraries in the "gmt" format.

For metabolomics, EnrichMiner supports *KEGG compounds* (https://www.genome.jp/kegg) and *SMPDB Primary* (https://smpdb.ca) two metabolite set libraries, which contain 81 and 104 terms respectively. Due to the inconsistency of metabolites nomenclature in published metabolite libraries, we recommend using InChI Key, CAS or KEGG ID of metabolites as input, or executing name matching via our in-home

function *Metabolite Name Fuzzy Matching*.

4.3 Enrichment Analysis Statistical Calculation

EnrichMiner supports the two mainstream enrichment analysis methods: ORA and GSEA. For ORA analysis, we use hypergeometric test. Raw p-values are adjusted via Benjamini-Hochberg method. The codes are based on Python *scipy* and *statsmodels* packages. In addition, EnrichMiner support ORA analysis provided by Enrichr (https://maayanlab.cloud/Enrichr). For GSEA, we use R package fgsea.

4.4 Pairwise Similarity calculation, network construction and community detection.

For each gene set library, we firstly pre-computed the pairwise similarities using python packages *numpy* and *scipy*. The clustering of the network is realized through the community detection algorithms, including Markov clustering, Informap method or Leiden method. Community detection function is implemented with python packages *igraph-python* and *markov_clustering*. Similarity network map is visualized with Cytoscape.js.

4.5 Directed acyclic graph

GO obo file is downloaded from gene ontology database. GOATOOLS is used to extract DAG of the terms of interest. DAG is visualized with Cytoscape.js.

4.6 Data availability

Gene lists and gene table referenced in this study are available in Supplementary Data.

4.7 Code availability

Custom code used in this study is available from the corresponding authors upon reasonable request.

Acknowledgment

This work has been supported by the National Key Research and Development Program of China (2021YFA1301604, 2021YFA1301603 and 2020YFE0202200), the National Natural Science Foundation of China (82100130, 32088101), and the CAMS Innovation Fund for Medical Sciences (CIFMS) (2019-I2M-5-063).

Author Contributions

L.X., XC.B and C.C. designed and co-supervised this project. L.X. and X.T. developed the frontend; L.X. and KK.X. developed the backend. C.C., L.X. and KK.X. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

References:

1. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).
2. Subramanian A*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

3. Geistlinger L*, et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform* **22**, 545-556 (2021).
4. Sherman BT*, et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* **50**, W216-W221 (2022).
5. Raudvere U*, et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191-W198 (2019).
6. Kuleshov MV*, et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97 (2016).
7. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* **47**, W199-W205 (2019).
8. Xia J, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* **38**, W71-77 (2010).
9. Wiredja DD, Koyuturk M, Chance MR. The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics* **33**, 3489-3491 (2017).
10. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600-1607 (2006).
11. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976-978 (2010).
12. Wang G, Oh DH, Dassanayake M. GOMCL: a toolkit to cluster, evaluate, and extract non-redundant associations of Gene Ontology-based functions. *BMC Bioinformatics* **21**, 139 (2020).
13. Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
14. Zhou Y*, et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
15. Liu X*, et al.* ExpressVis: a biologist-oriented interactive web server for exploring multi-omics data. *Nucleic Acids Res* **50**, W312-W321 (2022).
16. Gao Q*, et al.* Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *Cell* **179**, 561-577 e522 (2019).

Table 1. The feature summary of EnrichMiner compared with existing similar tools.

	EnrichMiner	WebGestalt	Enrichr	GOMCL	REVIGO	Metascape
Data						
Input data type	Genes of interest	Genes of interest	Genes of interest	Enriched terms	Enriched terms	Genes of interest
Functions						
ORA	√	√	√	-	-	√
GSEA	√	√	-	-	-	-
Redundancy deduction	√	√	-	√	√	√
Clustering of enriched terms	√	-	-	√	-	√
Interactive operations						
Interactive figures	√	partly	√	-	√	-
Interactive tables	√	partly	-	-	√	-
Visualizations						
DAG	√	√	-	√	-	-
Visualization for ORA/GSEA	√	√	-	-	-	partly

- EnrichMiner: https://omicsmining.ncpsb.org.cn/ExpressVis/EnrichMiner
- WebGestalt: http://www.webgestalt.org/
- Enrichr: https://maayanlab.cloud/Enrichr/
- GOMCL: https://github.com/Guannan-Wang/GOMCL
- REVIGO: http://revigo.irb.hr/
- Metascape: https://metascape.org/

Supplementary Information

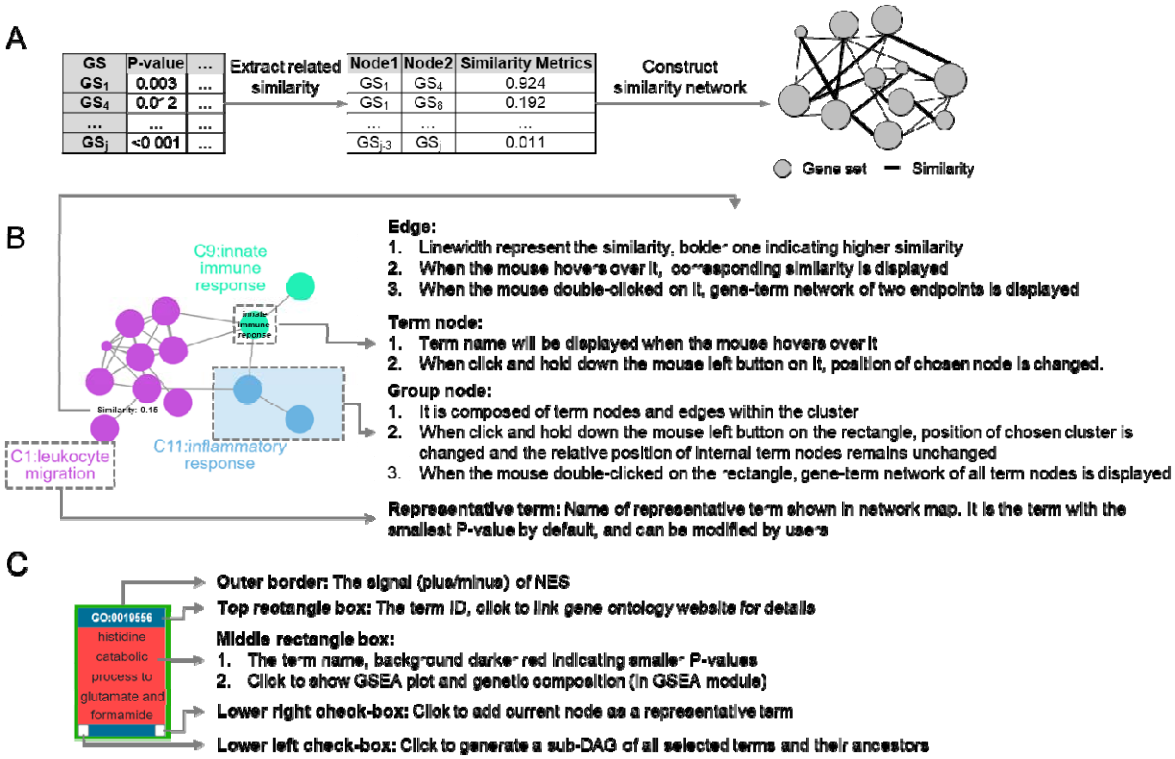

Figure S1. Details of reducing redundancy and selecting representative terms. (A) Processing to construct similarity network. For each gene set library, a data matrix of pairwise similarity is pre-calculated and saved as three columns (Node1, Node2 and Similarity Metrics). In each enrichment analysis task, similarities related to significantly enriched terms are extracted and used to construct similarity network. Nodes represent gene sets and edges represent similarity metrics. (B) Description of nodes and edges in network map. (C) Descriptions of five part in a DAG node (the outer border, the top rectangle box, the middle rectangle box, the lower left check-box and the lower right check-box).