CAMO R Shiny Tutorial

Contents

1 Introduction **1**
- 1.1 Abbreviation terms . . . 2

2 Preliminaries **2**
- 2.1 Citing CAMO . . . 2
- 2.2 How to start CAMO . . . 2
  - 2.2.1 Start from R . . . 2
  - 2.2.2 Run from docker image: . . . 3
- 2.3 CAMO setting page . . . 3
- 2.4 Question and bug report . . . 4

3 Preparation **4**
- 3.1 Data input . . . 4
  - 3.1.1 Preprocessed gene-expression data . . . 4
  - 3.1.2 Clinical data . . . 5
  - 3.1.3 Pre-calculated p-value data . . . 5
  - 3.1.4 Real data examples in the CAMO software . . . 5
- 3.2 Optional input for analyses . . . 6
  - 3.2.1 Ortholog matching file . . . 6
  - 3.2.2 Pathway database . . . 7
  - 3.2.3 Noun-pathway matrix . . . 7
  - 3.2.4 Gene annotation file . . . 7
  - 3.2.5 Pathway ID annotation file . . . 10

4 Data uploading, preprocessing and merging **12**
- 4.1 Data uploading and preprocessing for each study . . . 12
- 4.2 Merging multiple studies at the 'Saved Data' page . . . 14

5 Analysis **15**
- 5.1 Genome-wide analysis . . . 15
- 5.2 Pathway-based analysis . . . 17
- 5.3 Individual pathway analysis . . . 23

6 Complete list of options **29**
- 6.1 Data Uploading and Preprocessing . . . 29
- 6.2 Saved Data . . . 29
- 6.3 Analysis: Genome-wide analysis . . . 30
- 6.4 Analysis: Pathway-based analysis . . . 30
- 6.5 Analysis: Individual pathway analysis . . . 31

1 Introduction

CAMO is an analytical software with R Shiny based graphical user interface (GUI) for evaluating omics congruence of model organisms. It performs threshold-free Bayesian differential analysis and generates quantitative concordance and discordance scores (c-scores and d-scores) both genome-wide and at pathway level

for all pair-wise studies. Based on the c-scores/d-scores, CAMO conducts a series of downstream machine learning and bioinformatics analysis with interactive visualization for pathway knowledge retrieval and topological gene module detection. Outputs from the tool will provide the foundation for hypothesis generation and subsequent translational investigation.

In this tutorial, we will go through the installation and usage of CAMO step by step using real data examples. The CAMO Shiny software and associated R package are available at `https://github.com/CAMO-R`. This tutorial can be found at `https://github.com/CAMO-R/other/tree/main/Rshiny_tutorial`.

1.1 Abbreviation terms

- General terms:
  - DE: Differentially expression
  - FC: Fold change
  - MDS: Multidimensional scaling
- Methods or tools:
  - MCMC: Markov Chain Monte Carlo
  - LIMMA: R package for the analysis of gene expression data arising from microarray or RNA-Seq technologies
  - DESeq2: R package for moderated estimation of fold change and dispersion for RNA-seq data

2 Preliminaries

2.1 Citing CAMO

Please cite appropriate papers if you use CAMO, by which the authors will receive professional credits for their work.

- Zong, W., Rahman, M. T., Zhu, L., Zeng, X., Zhang, Y., Zou, J., Liu, S., Ren, Z., Li, J. J., Oesterreich, S., et al. (2021). Camo: A molecular congruence analysis framework for evaluating model organisms. *bioRxiv*

2.2 How to start CAMO

2.2.1 Start from R

Requirement:

- R >= 4.0.0
- Rcpp >= 1.0.0
- Shiny >= 1.0.0
- C++11 Compiler

Note:

- We recommend users to use R 4.0 to implement our tool. If you are using R 3.5 or lower, you may encounter errors in installing dependencies of the modules. You can manually install the dependencies.
- MacOS users may need to install xcode first. In terminal, run following:
  $x$code-select –install
- MacOS users may encounter error "*** C++11 compiler required; enable C++11 mode in your compiler, or use an earlier version of Armadillo", which can be solved by run following code command in R Console:
  Sys.setenv("PKG_CXXFLAGS"="-std=c++11")

How to install the software:

1. Install the CAMO R package and dependency packages following the instruction at `https://github.com/CAMO-R/Rpackage`.
2. Download the CAMO Shiny project at `https://github.com/CAMO-R/Rshiny` by clicking on "code > Download ZIP" and extract to a local folder named as "Rshiny".

How to start the software:

1. Open "RunShiny.R" file in R console.
2. Set the working directory of R to the directory "path_to_Rshiny/" which contains the Shiny project folder "Rshiny" (Figure 1 (1)).
3. Click on the "Run App" button (Figure 1 (2)).

   Note that the installation progress of R packages may take up to a few minutes. Users can check the progress in R console and may need to select whether to update all/some/none packages. After all packages has been installed, the CAMO Shiny app will automatically open in your default browser.

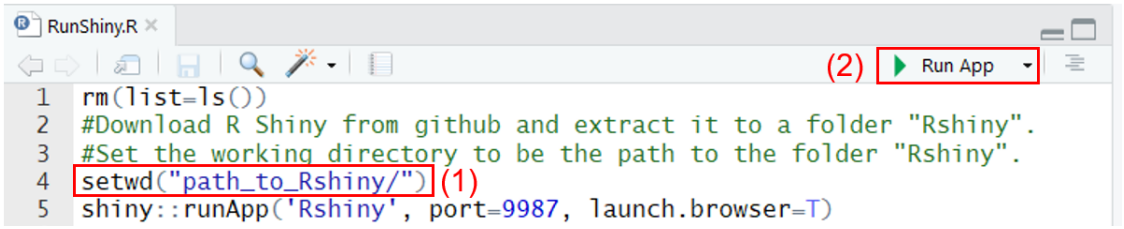

Figure 1: Running CAMO App

2.2.2 Run from docker image:

1. Install docker desktop(`https://docs.docker.com/desktop/`) following the installation instruction for your operating system.
2. In terminal:

   docker pull weiiizong/camo

   docker run –rm –name camo -p 3838:3838 weiiizong/camo:1.0
3. Go to your web browser on `http://127.0.0.1:3838/CAMO/`.

2.3 CAMO setting page

After starting CAMO, the first page is the CAMO setting page as shown in Figure 2. There are four tabs on the top of this page (see Figure 2 (1)), which will direct users to specific functional modules of the software including "Setting," "Data Uploading and Preprocessing," "Saved Data," and "Analysis". Below these tabs is a **Welcome to CAMO** page, which briefly introduces the software and other information about the authors and maintainers. **Session Information** summarizes the current server information. **Directory for Saving Output Files** (see Figure 2 (2)) allows users to select a working directory to save all results generated during the computation. This has to be set before any analysis. The current working directory is displayed on the top-right corner (see Figure 2 (3)). Note that some tables and figures displayed on the user interface are dependent on the results in the working directory. For example, if the working directory already contains the genome-wide c-scores and d-scores calculation results (i.e., "ACS_ADS_Global.RData"), the table will automatically been demonstrated on the Genome-wide analysis panel. Please make sure the working directory contains the results if any intended to show and always change to a new working directory to initiate new congruence analysis.

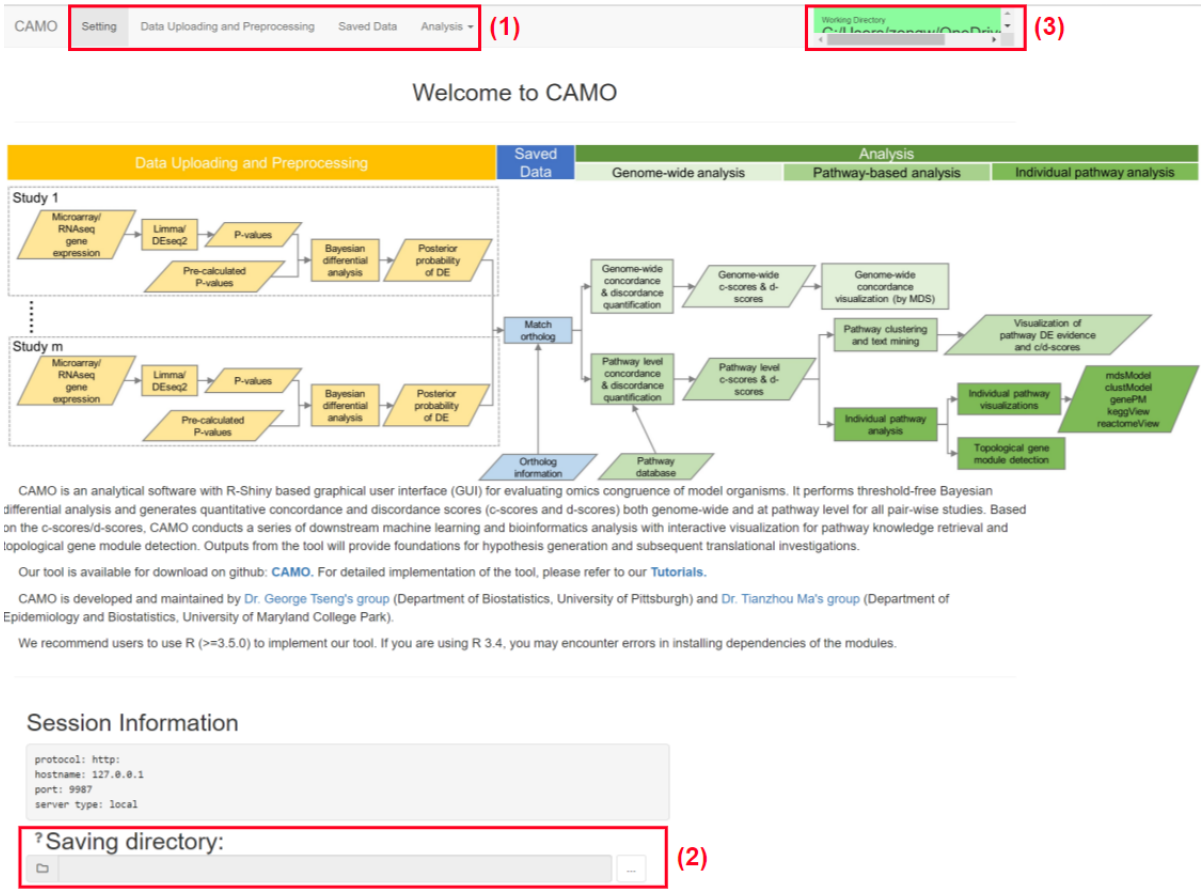

Figure 2: CAMO software suite GUI setting page

2.4 Question and bug report

If you encounter errors or bugs, please report at `https://github.com/CAMO-R/Rshiny/issues`.

3 Preparation

In this section, we will introduce how to prepare input for CAMO including both the required study data and the optional input (e.g., pathway database, noun-pathway matrix etc.) for downstream analyses.

3.1 Data input

CAMO software allows two types of input data: a preprocessed gene-expression data matrix file with corresponding clinical data file, or a pre-calculated p-value data matrix file from DE analysis.

3.1.1 Preprocessed gene-expression data

The gene-expression matrix should be prepared as a comma-separated ".csv" file (see an example in Figure 3). The first column (row names) is the feature ID (e.g., gene symbol, probe ID, or entrez ID) and the rest of columns are the expression data for samples whose IDs are annotated as column names. Valid data types include microarray data and RNA-seq count data. Note that each dataset needs to be carefully preprocessed following the standard pipeline (filtering, normalization, removal of redundant probes for microarray, removal of low counts genes for RNA-Seq, annotation, etc.) and ready for standard DE analysis such as LIMMA or DEseq2.

	A	B	C	D	E	F	G	H	I
1		GSM178612	GSM178613	GSM178614	GSM178615	GSM178628	GSM178629	GSM178630	GSM178631
2	Zglp1	3.738	3.107	4.212	4.282	3.993	3.832	3.802	3.927
3	Oog3	3.454	3.517	3.792	3.404	3.277	3.472	3.459	3.366
4	Ldlrap1	5.858	5.595	5.014	6.06	5.795	6.222	5.832	6.103
5	Mdn1	6.758	6.111	5.826	6.914	7.325	7.156	7.356	7.489
6	Wfdc17	11.254	11.671	11.199	11.429	10.723	10.027	11.002	9.726
7	9330171B	2.642	2.775	2.779	2.684	2.612	2.712	2.721	2.661
8	1700024h	3.161	3.363	3.174	3.203	3.152	3.2	3.116	3.422
9	2610305J2	3.46	3.328	3.041	3.548	3.593	3.239	3.327	3.168
10	A930017K	3.162	3.109	3	2.801	2.941	2.877	2.778	2.834

Figure 3: A gene-expression data format example

3.1.2 Clinical data

Clinical data should be prepared as a comma-separated ".csv" file (see an example in Figure 4). It has two columns, the first column (row names) can be arbitrary, but the second column ("group") contains the group labels (e.g., case/control, treated/non-treated) of the samples in the gene expression data. The group labels should in the same order as the sample IDs on the column of the gene expression data to avoid any mismatch issues. The second column should be named by "group" as shown in Figure 4.

	A	B
1		group
2	1	2
3	2	2
4	3	2
5	4	2
6	5	1
7	6	1
8	7	1
9	8	1

Figure 4: A clinical data format example.

	A	B	C
1		pvalue	logFC
2	Zglp1	0.834152	-0.05375
3	Oog3	0.273289	0.14825
4	Ldlrap1	0.157273	-0.35625
5	Mdn1	0.004534	-0.92925
6	Wfdc17	0.005927	1.01875
7	9330171B17Rik	0.683326	0.0435
8	1700024h08rik	0.98211	0.00275
9	2610305J24Rik	0.937499	0.0125
10	A930017K11Rik	0.218934	0.1605

Figure 5: A pre-calculated p-value data example

3.1.3 Pre-calculated p-value data

The pre-calculated p-value data should be prepared as a comma-separated ".csv" file generated from DE analysis comparing the two groups.(see example in Figure 5). It has three columns, the first column (row names) is the feature ID (e.g., gene symbol, probe ID, or entrez ID), the second column (named as "pvalue") and the third column (named as "logFC") are the p-values and logFC values of the corresponding features. The first row contains the column names and should indicates the "pvalue" column and "logFC" column.

3.1.4 Real data examples in the CAMO software

We collected two multi-cohort cross-species datasets for the CAMO software, including the 12 inflammatory models in human and mouse (HM) example and the 10 developmental stage models in C. elegans (CE) and D. melanogaster (DM) example. The Bayesian differential analysis results for each individual data have been saved to the CAMO Shiny app whose meta data can be seen at the "Saved Data" page, which can be used for merging and analysis directly. The data spreadsheets are provided at `https://github.com/CAMO-R/other/tree/main/Example_data`.

Table 1 summarizes the studies in the human-mouse inflammatory models example. Table 2 summarizes the studies in the CE-DM developmental stages example. The HM data is used for demonstration in this tutorial.

Table 1: A total of 12 microarray inflammatory response studies in human and mouse including six inflammatory response studies in human (Burns, Infection, Trauma , Sepsis, Endotoxin (LPS) and Acute Respiratory Distress Syndrome (ARDS), abbreviated as HB, HI, HT, HS, HL and HA) and six corresponding studies in mouse (abbreviated as MB, MI, MT, MS, ML and MA). The 12 microarray datasets in Affymetrix and Illumina platforms are preprocessed and normalized as uniformly as possible. When multiple time points are measured, we evaluate the cross-species time series and selected the best matched time points in transcriptomic response (i.e, the time point with largest c-score or d-score) as representatives. Two groups of data labeled as 1 and 2 corresponds to control and case samples.

Species	Condition	# Genes	# Samples (control/case)
human	Burn	20107	37/21
human	Infection	25160	44/16
human	Trauma	20107	44/16
human	Sepsis	20107	37/23
human	LPS	12437	22/4
human	ARDS	12493	21/13
mouse	Burn	20672	4/4
mouse	Infection	13015	17/11
mouse	Trauma	20672	4/4
mouse	Sepsis	20672	5/5
mouse	LPS	20672	4/4
mouse	ARDS	22039	2/3

Table 2: A total of 35 worm samples measured at four developmental stages (embryo, larvae, dauer and adult) and 30 fruit fly samples measured at four developmental stages (embryo, larvae, pupae and adult). The processed log2-transformed RNA-seq FPKM (Fragments Per Kilobase of transcript per Million mapped reads) data are from Li et al. (2014). Data are further splited to 5 comparisons in each species (10 comparisons in total) by treating worm adult and fruit fly female adult as the reference group, i.e., CE: early embryo vs. adult, mid embryo vs. adult, late embryo vs. adult, larvae vs. adult, dauer vs. adult; DM: early embryo vs. female adult, mid embryo vs. female adult, late embryo vs. female adult, larvae vs. female adult, pupae vs. female adult.

Species	Phase	# Genes	# Samples (adult/non-adult)
CE	early embryo	19166	2/9
CE	mid embryo	19166	2/6
CE	late embryo	19166	2/9
CE	larvae	19166	2/6
CE	dauer	19166	2/3
DM	early embryo	13028	3/5
DM	mid embryo	13028	3/3
DM	late embryo	13028	3/4
DM	larvae	13028	3/6
DM	pupae	13028	3/6

3.2 Optional input for analyses

3.2.1 Ortholog matching file

An ortholog matching file is required for matching and merging studies at the “Saved Data” page. 5 popular ortholog matching files are provided by CAMO (see Figure 6). In addition, CAMO allows users to upload their own ortholog matching file which should be a dataframe or matrix object in R environment and saved as a ".RData" or ".rda" file. The dataframe or matrix should have two columns matching orthologs between two species with their species names as the column names. Figure 7 is an example of a matrix matching orthologs between homo sapiens (hs) and mus musculus (mm) in R environment.

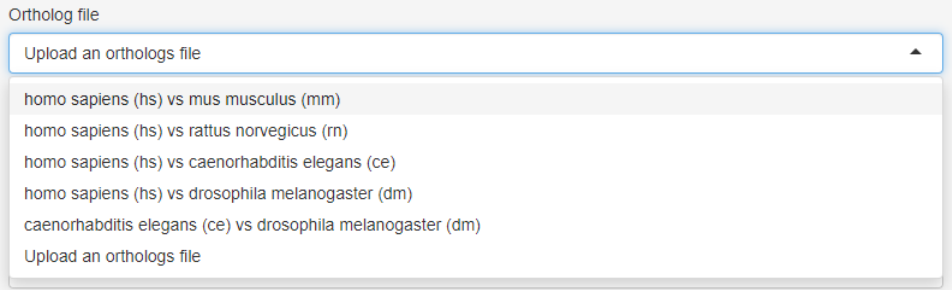

Figure 6: Options for the ortholog matching file in CAMO.

	hs	mm
1	SPINK7	Spink7
2	ACBD7	Dbil5
3	NEGR1	Negr1
4	POMGNT2	C85492
5	CCR8	Ccr8
6	CHRFAM7A	Chrna7
7	COPS3	Cops3
8	CCDC71	Ccdc71
9	ABCG4	Abcg4
10	TUSC2	Tusc2

Figure 7: A example of a matrix matching orthologs between homo sapiens (hs), mus musculus (mm) in R environment

3.2.2 Pathway database

Popular databases for multiple sepecies are provided in the CAMO software (see Figure 8) for pathway-level c-scores/d-scores calculation. However, users are allowed to upload their own pathway gene sets if they have other preference. The pathway gene sets should be prepared as a list object in R environment and saved as a ".RData/.rda" file containing all pathways intended to be used in the analyses. Each pathway should be an element in the list named by the pathway name and valued by a character vector of genes it contains. Figure 9 is an example of the pathway list format in R environment.

3.2.3 Noun-pathway matrix

A flat noun-pathway matrix is used in the text mining step. Please refer to Zeng et al. (2020) for details. CAMO provides noun-pathway matrices generated from KEGG and Reactome pathway descriptions for homo sapiens (hsa), mus musculus (mmu), rattus norvegicus (rno), caenorhabditis elegans (cel) and drosophila melanogaster (dme) (see Figure 10). It also allow user to upload their own noun-pathway matrix if they have other preference. The noun-pathway matrices should be prepared as a dataframe or a matrix object in R environment and saved as a ".RData/.rda" file. It should have four columns in the order of "phrase", "row", "col", "value" and "pathway". The "row" and "col" are the location indices of the corresponding "phrase" and "pathway" in the large matrix and "value" is the penalized sum of the number of times it appears in the pathway description. Figure 11 is a noun-pathway matrix example in R environment.

3.2.4 Gene annotation file

To utilize the topological visualization (for KEGG and Reacotme pathways) and the topological module detection function (for KEGG pathways). The gene names in data and topological plots should be the same type.

For the topological visualization of KEGG pathways, the R package *pathview* is called to annotate gene nodes on the KEGG topological plots which requires the genes to be named by their entrez IDs. If the gene

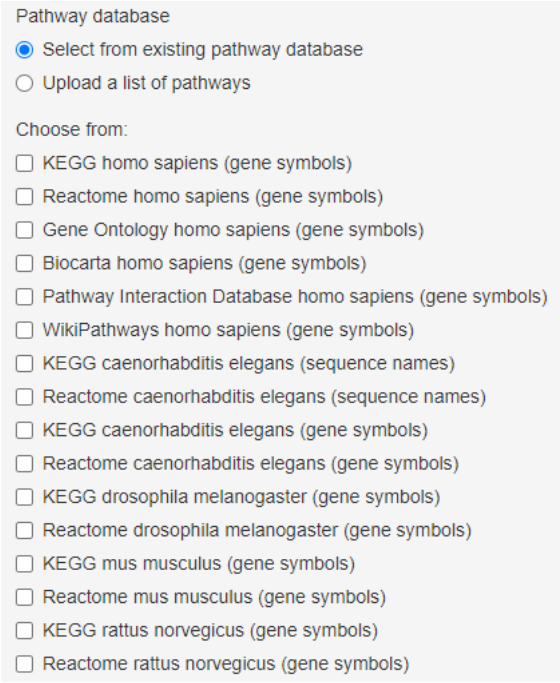

Figure 8: Pathway databases included in CAMO.

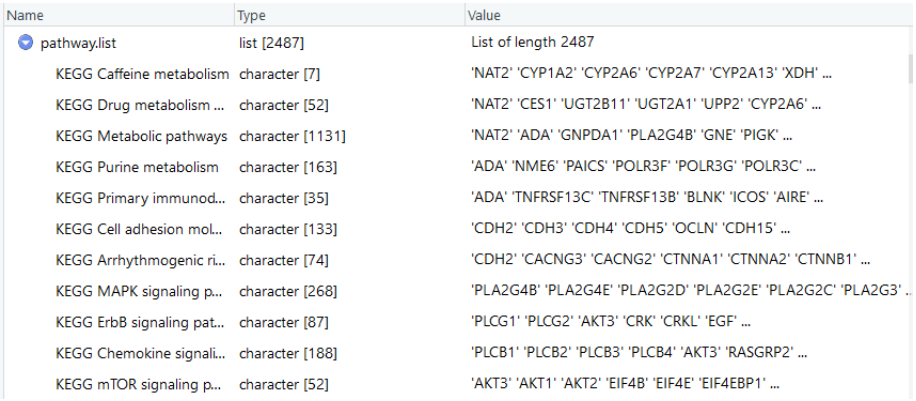

Figure 9: A pathway list example in R environment.

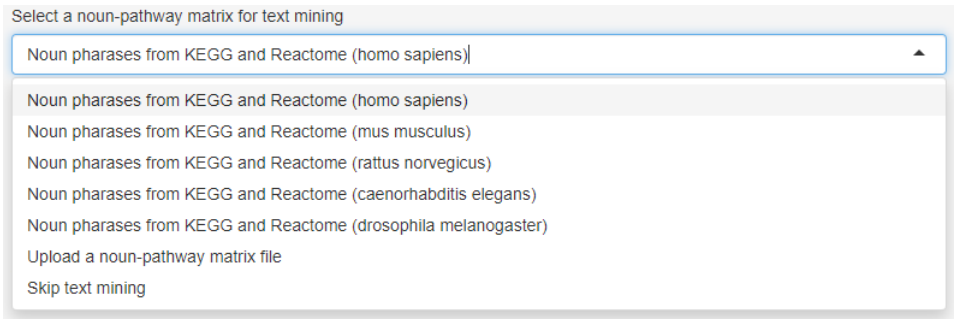

Figure 10: Options for noun-pathway matrix in CAMO.

	phrase	row	col	value	pathway
acetyl.coa	acetyl coa	11	1	0.22313016	KEGG Glycolysis / Gluconeogenesis
acid.metabolism	acid metabolism	14	1	1.00000000	KEGG Glycolysis / Gluconeogenesis
biosynthesis	biosynthesis	93	1	0.22313016	KEGG Glycolysis / Gluconeogenesis
diagram	diagram	223	1	0.22313016	KEGG Glycolysis / Gluconeogenesis
metabolite	metabolite	491	1	0.22313016	KEGG Glycolysis / Gluconeogenesis
pyruvate	pyruvate	651	1	0.22313016	KEGG Glycolysis / Gluconeogenesis

Figure 11: A noun-pathway matrix example in R environment.

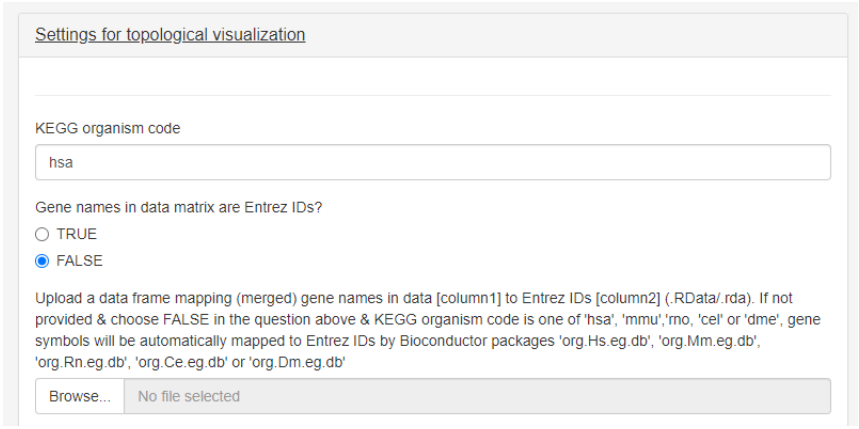

Figure 12: Options for gene annotation file for KEGG topological visualization in CAMO

names in the input data is not in entrez ID, CAMO provides automatic annotation by Bioconductor packages (see Figure 12). Alternatively, users can provide their own annotation file by uploading an annotation dataframe or matrix saved as ".RData/.rda" file. The dataframe should have two columns where the first column is the gene names in data and the second column contains their corresponding entrez IDs. Figure 13 is a user-defined gene annotation matrix example in R environment.

For the topological visualization of Reactome pathways, we developed our own tool to first retrieve and parse the pathway topology from Reactome database using the Python minidom parser and then color each node by its concordance/discordance information using the Python Imaging Library. It assumes that the data gene names are consistent with gene names on the Reactome topological figures downloaded from the Reactome website which are usually gene symbols. User can provide their own annotation file by uploading an annotation dataframe or matrix saved as ".RData/.rda" file of the same format as in Figure 13. The only difference is that instead of entrezID, the second column should be the gene names used on Reactome topological figures.

For the topological module detection function of KEGG pathways, we retrieved the gene node/entry connections on the KEGG topological plots by downloading their .xml files from KEGG database using the R package *KEGGgraph* and construct graph from it. Therefore, the data gene names should be consistent with the gene node/entry names stored in the .xml files. The entry name type can be different for different species which can be checked by clicking nodes on the interactive topological figures on the KEGG website. For homo sapiens (hsa), mus musculus (mmu), rattus norvegicus (rno), caenorhabditis elegans (cel) and drosophila melanogaster (dme), CAMO fills this gap by mapping between gene symbols and the entry names appeared in KEGG topological figures. However, advanced users can provide their own annotation file by uploading a ".RData/.rda" file containining an annotation dataframe or matrix of the same format as in Figure 13. The only difference is that instead of entrezID, the second column should be the entry names used in KEGG topological figures (the .xml file on KEGG website).

	gene	entrezID
A1B	A1B	1
ABG	ABG	1
GAB	GAB	1
HYST2477	HYST2477	1
A1BG	A1BG	1
A2MD	A2MD	2
CPAMD5	CPAMD5	2
FWP007	FWP007	2
S863-7	S863-7	2
A2M	A2M	2
A2MP	A2MP	3
A2MP1	A2MP1	3

Figure 13: A gene annotation matrix example in R environment

3.2.5 Pathway ID annotation file

Due to the complex interactions across platforms and packages in the topological visualization and module detection. Pathway IDs can be provided for each KEGG and Reactome pathway name to avoid confusion by uploading a pathway name list named after IDs (".RData/.rda" file). An example of the pathway annotation list in R environment is shown in Figure 14. If not provided, CAMO will automatically match each path name by the R package *KEGGREST* for KEGG pathways or *reactome.db* for Reactome pathways.

Name	Type	Value
KEGG.pathID2name	list [345]	List of length 345
hsa00010	character [1]	'Glycolysis / Gluconeogenesis'
hsa00020	character [1]	'Citrate cycle (TCA cycle)'
hsa00030	character [1]	'Pentose phosphate pathway'
hsa00040	character [1]	'Pentose and glucuronate interconversions'
hsa00051	character [1]	'Fructose and mannose metabolism'
hsa00052	character [1]	'Galactose metabolism'
hsa00053	character [1]	'Ascorbate and aldarate metabolism'
hsa00061	character [1]	'Fatty acid biosynthesis'
hsa00062	character [1]	'Fatty acid elongation'
hsa00071	character [1]	'Fatty acid degradation'
hsa00100	character [1]	'Steroid biosynthesis'
hsa00120	character [1]	'Primary bile acid biosynthesis'
hsa00130	character [1]	'Ubiquinone and other terpenoid-quinone biosynthesis'
hsa00140	character [1]	'Steroid hormone biosynthesis'
hsa00190	character [1]	'Oxidative phosphorylation'
hsa00220	character [1]	'Arginine biosynthesis'
hsa00230	character [1]	'Purine metabolism'
hsa00232	character [1]	'Caffeine metabolism'
hsa00240	character [1]	'Pyrimidine metabolism'
hsa00250	character [1]	'Alanine, aspartate and glutamate metabolism'
hsa00260	character [1]	'Glycine, serine and threonine metabolism'

Figure 14: An pathway annotation list example in R environment.

4 Data uploading, preprocessing and merging

In this section, we introduce how to upload, preprocess single study data and merge them by ortholog matching file before c-scores/d-scores calculation.

4.1 Data uploading and preprocessing for each study

Step 1 Upload and preprocess data:

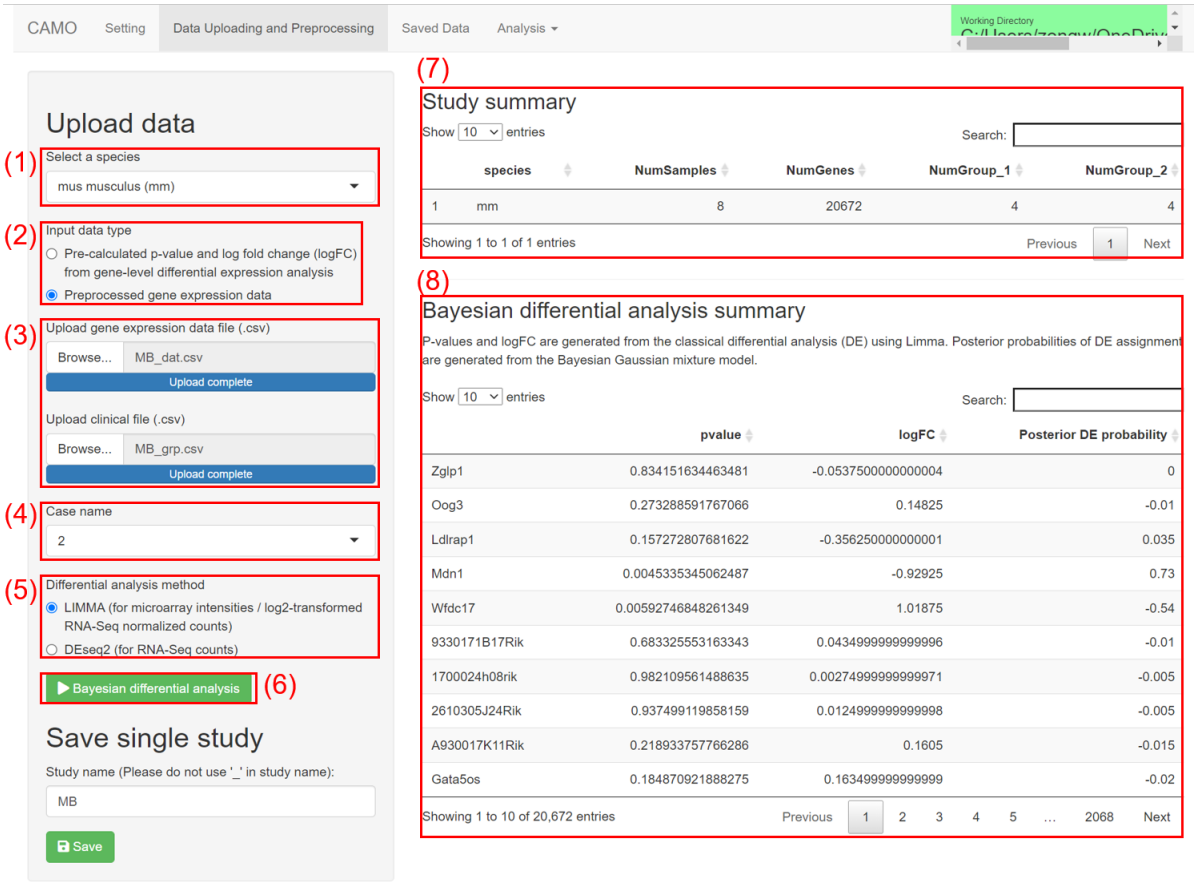

Figure 15: GUI for data uploading and preprocessing

For each study, the user needs to first specify the species of the study. Five species are provided for users to choose from: homo sapiens (hs), mus musculus (mm), rattus norvegicus (rn), caenorhabditis elegans (ce) and drosophila melanogaster (dm). If none of them applied, the user can also select the "other" option to type in the species name at the "Select a species" box (see Figure 15 (1)).

Then, users can specify the "Input data type" (see Figure 15 (2)) based on the types of their prepared data files which should be prepared according to Section 3.1. By clicking the "Browse..." buttons (see Figure 15 (3)), users can select the file to be uploaded from local directory. For preprocessed gene expression data, users should specify the case label at the "Case Name" box (see Figure 15 (4)) which determines the comparison direction within each study. This should be as consistent as possible across studies for the following concordance analysis to be meaningful. In addition, users need to select the DE analysis used to derive p-values and logFC as an input to the threshold-free Bayesian differential analysis (see Figure 15 (5)). Two options are allowed, the LIMMA is suitable for microarray intensities or log2-transformed RNA-seq normalized counts which is more Normally distributed while DEseq2 is more suitable for counts data.

After the data been successfully uploaded, users can click the "Bayesian differential analysis" button (see Figure 15 (6)) at the side panel to obtain the posterior DE probability of each gene. If the data type is prepeocessed gene expression data, classical DE analysis (LIMMA/DEseq2) will be performed first to generate p-values and logFC for each gene. Then, the Bayesian differential analysis is applied on the one-sided p-values to derive posterior DE probabilities. Since this step has a MCMC procedure, it may take a while to run. After the calculation been completed, a study summary table (see Figure 15 (7)) and a Bayesian differential analysis summary table (see Figure 15 (8)) will show up to the right. The Bayesian differential analysis summary table contains the raw p-values and logFC from the classical DE analysis and the posterior DE probability (last column) from the Bayesian modelling.

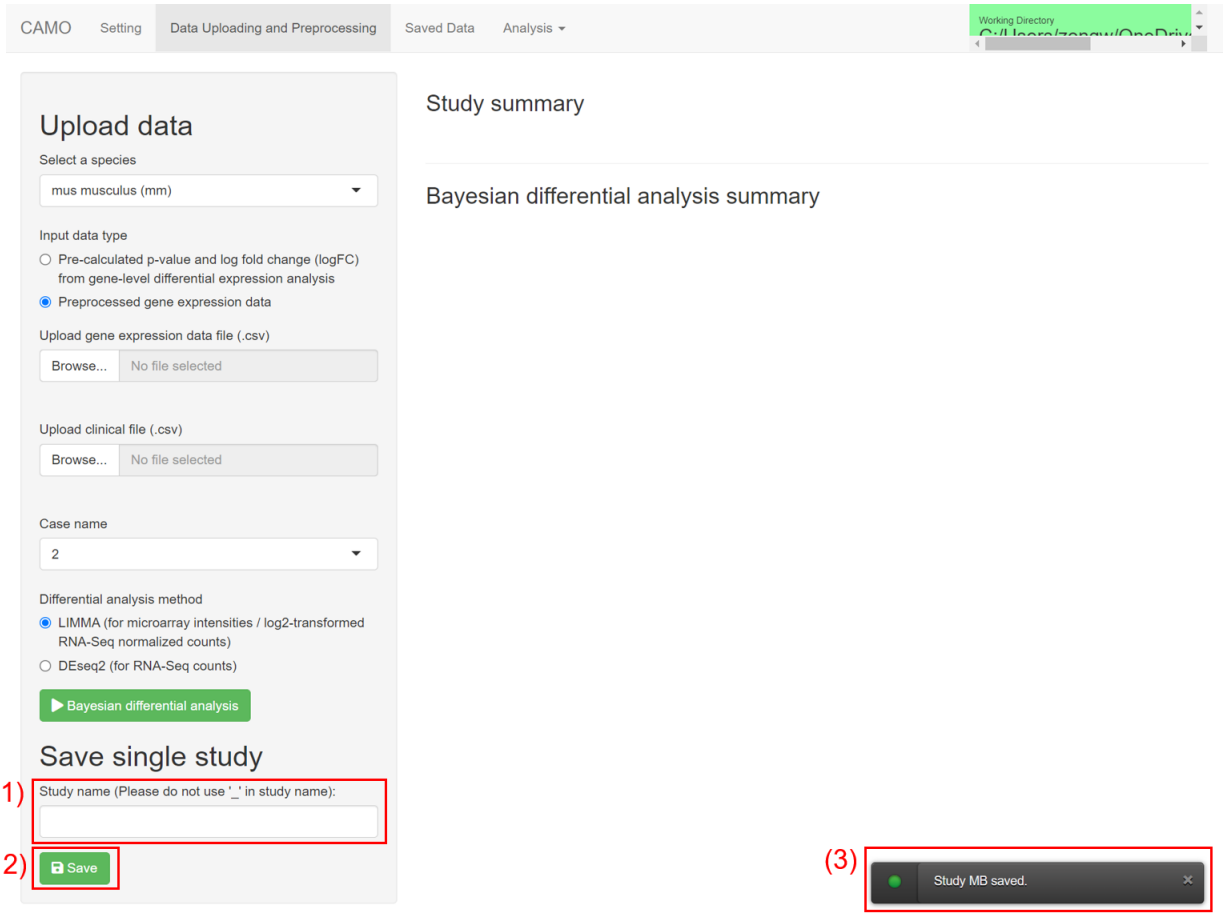

Figure 16: GUI for saving a study.

Step 2 Save results as a single study: In the next step, users can type a study name at the "Study name (Please do not use '_' in study name):" box (see Figure 16 (1)), and click the "Save" button (see Figure 16 (2)) to save the results. Note that users should not include symbol "_" in the study names to avoid confusion in the downstream analysis. Then, a message will pop up at the right bottom corner (see Figure 16 (3)) if it is successfully saved in the database and can be used for further analyses. The study name should be unique for each study. After each study been saved, the web page will be cleared and users can repeat the steps above until all studies of interests are saved successfully. All saved studies are then available at the "Saved Data" page (see Section 4.2) and ready for the following analyses.

4.2 Merging multiple studies at the 'Saved Data' page

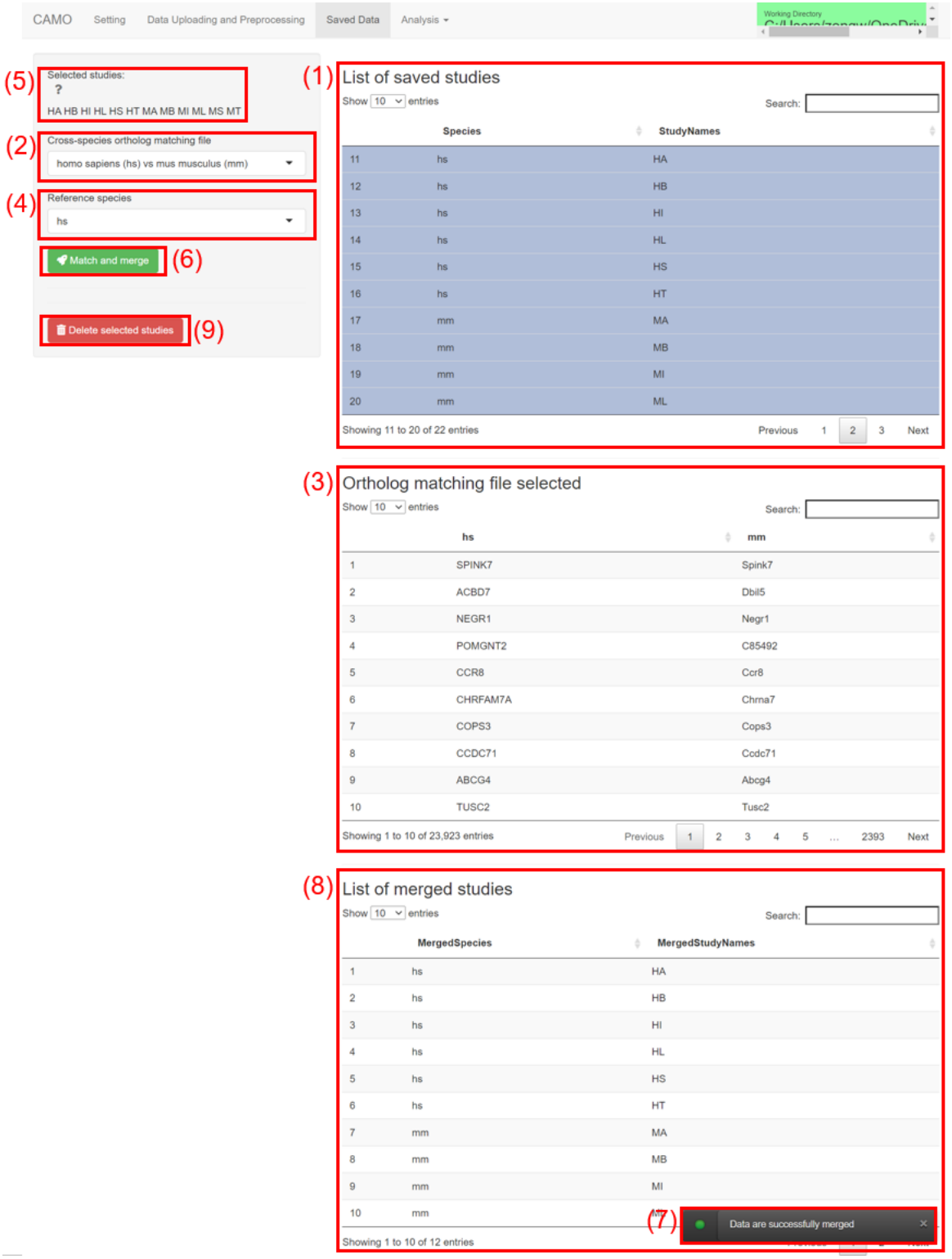

Figure 17: GUI for merging saved data.

After uploading at least one study for each of the two species, users can move to the merging step at the Saved Data page. All previous saved data are summarized at the "List of saved studies" (see Figure 17

(1)). For same species comparison, please select "Same species - skip orthologs matching" at Figure 17 (2) to skip orholog matching. For cross-species comparison, an ortholog matching file is required for matching and merging studies. It can be selected from existing ones or uploaded from a local directory (see Figure 17 (2) and Section 3.2.1 for ortholog file preparation). The content of selected ortholog file can be viewed at "Ortholog matching file selected" on the right panel (see Figure 17 (3)) after been selected or uploaded at the "Cross-species ortholog matching file" box on the left side panel. One of the species should be selected as the reference species (see Figure 17 (4)). Genes from the other species will be matched to genes in this reference species by the ortholog matching file selected.

Users can select the studies of interests by clicking on the rows of the "List of saved studies" to merge. Selected study names will be shown at the top of the left panel (see Figure 17 (5)). After selecting the studies and clicking on the "Match and Merge" button (see Figure 17 (6)), the merging step will start which takes about a minute to run. If this process has been done successfully, a message will pop up at the right corner (see Figure 17 (7)) and merged studies will be summarized at "List of merged studies" (see Figure 17 (8)).

Saved datasets can be deleted by first clicking on the rows of the "List of saved studies" to select and then clicking the "Delete selected datasets" (see Figure 17 (9)). Please be careful with deletion since it is permanent and not retrievable.

5 Analysis

After all studies are processed according to Section 4, the merged studies are ready for the analytical modules in CAMO. By clicking on the "Analysis" tab, users can navigate to "Genome-wide analysis", "Pathway-based analysis" or "Individual pathway analysis" to conduct genome-wide and pathway-level c-scores & d-scores and perform hypothesis generation and subsequent translational investigations based on the scores. In the next few subsections, we will introduce how to run each of the modules in detail.

Note that some analysis steps are computationally demanding and may take minutes or hours, depending on the data size, users can keep track of the progress by checking the R console.

5.1 Genome-wide analysis

By clicking on the "Genome-wide analysis" under the "Analysis" tab, users are directed to the genome-wide congruence analysis page.

There are two main functions ("Genome-wide c-scores & d-scores calculation" and "Genome-wide MDS plot") on the left side panel and two corresponding tabs on the right main panel. These two main analyses at genome-wide level should be done in order(i.e, genome-wide c-scores and d-scores should be calculated before the MDS map).

1. **Genome-wide c-scores & d-scores calculation:**

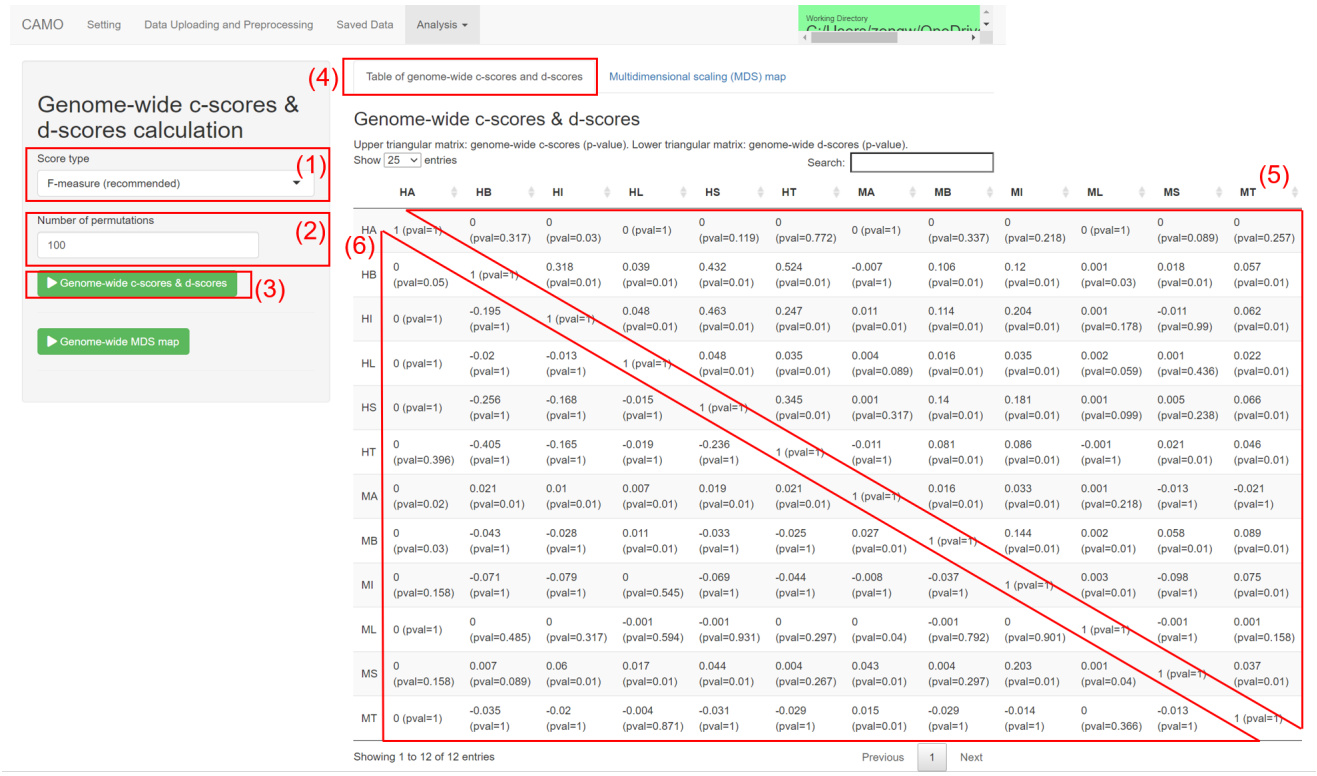

Figure 18: GUI for genome-wide c-scores and d-scores calculation.

Two parameters need to be selected at the left side panel before starting the c-scores & d-scores calculation.

- "Score type" (see Figure 18 (1)). Three types of congruence scores defined in the CAMO paper are F-measure, Youden index and geometric mean. The definition of each score type is described in detail in the CAMO paper. F-measure is recommended because it is symmetric i.e., same value for a pair regardless of which study is served as the reference group to be compared with.
- "Number of permutation" (see Figure 18 (2)) is the number of permutations during calculating the scores. 100 is used in our real data evaluation.

If users are not sure about parameter selection, please leave them as default which generates good result in general based on our real data evaluation. A complete list of options is also available in Section 6.3.

Then users can click on the "Genome-wide c-scores & d-scores" button (see Figure 18 (3)) to start c-scores and d-scores calculation. This step may take a while because of the permutation. Users can check the process at the R console. Once the calculations is finished, a summary table of the c-scores and d-scores and their corresponding p-values for all study pairs are shown at the "Table of genome-wide c-scores d-scores" tab (see Figure 18 (4)) to the right. The closer the c/d-score is to 1, the more concordant/discordant the two studies are and vice versa. A c/d-score near to 0 indicates close to randomly generated values and less than 0 being even worse. The upper triangular matrix (see Figure 18 (5)) contains the c-score (p-value) of of corresponding study pairs (row vs column) and the lower triangular matrix (see Figure 18 (6)) contains the d-score (p-value) of corresponding study pairs (row vs column).

The summary table displayed in Figure 18 is generated from the human-mouse inflammatory models using default parameters. Genome-wide c-socres and d-scores are saved as "ACS_ADS_global.RData" at the saving directory and a subdirectory "ACS_ADS_Global" is generated to store the c-socres and

d-scores tables as .csv files for easy investigation.

2. **Genome-wide MDS map:**

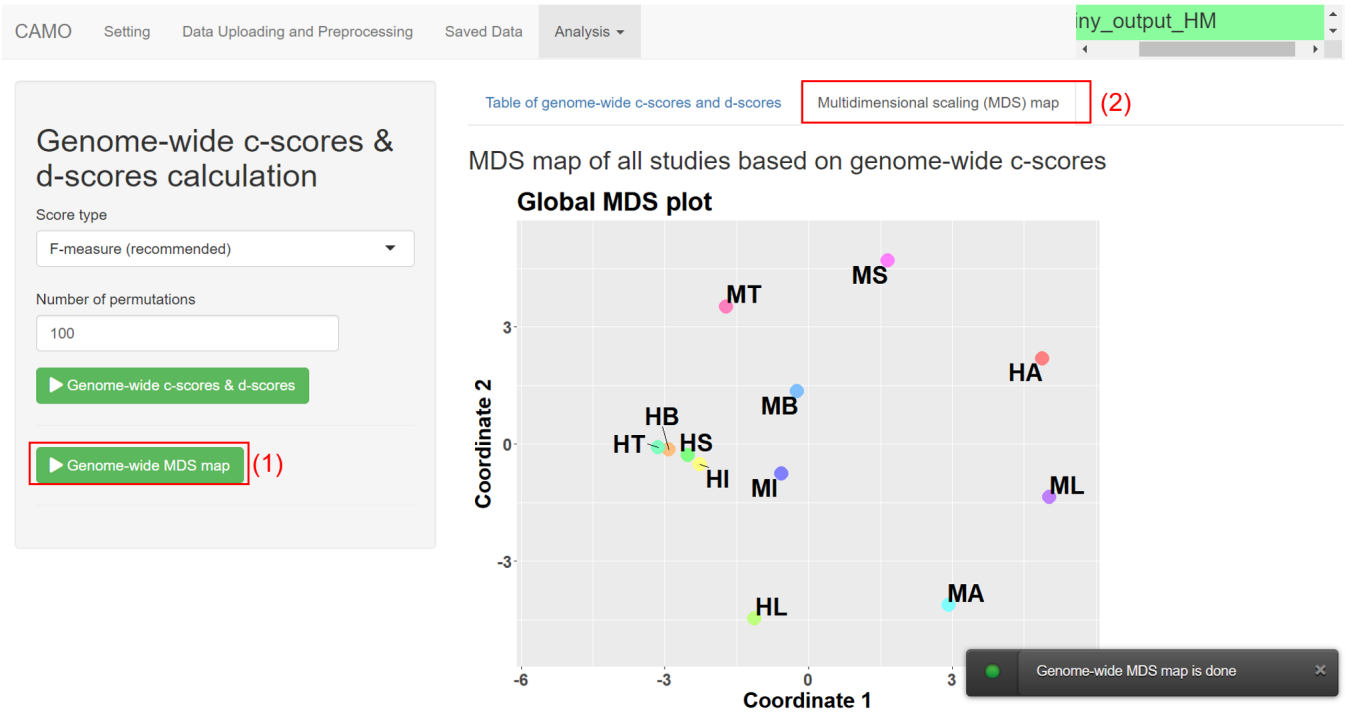

Figure 19: GUI for genome-wide MDS.

After the c-scores and d-scores been calculated, a MDS map can be used to visualize all studies based on the pairwise dissimilarity measure defined from the transformation of c-scores. After clicking on the "Genome-wide MDS map" button (see Figure 19 (1)), the MDS map will be displayed at the "Multidimensional scaling (MDS) map" at the right main panel (see Figure 19 (2)). This MDS file is saved as "global_MDS.pdf" at the saving directory.

Figure 19 shows the genome-wide MDS map for the human-mouse inflammatory model example.

5.2 Pathway-based analysis

By clicking on the "Pathway-based analysis" under the "Analysis" tab, users are directed to the pathway-level congruence analysis page.

There are three main functions ("Pathway c-scores & d-scores calculation", "Pathway clustering" and "Visualization of pathway DE evidence and c/d-scores") on the left side panel. The five tabs on the right main panel ("Table of pathway-level c-scores & d-scores", "Scree plot", "Visualization of pathway clusters", "Text mining on each pathway cluster" and "DE evidence and c/d-scores plot") will demonstrate the results generated from these three functions. The three main analyses should be done in order(i.e, pathway-level c-scores and d-scores should be calculated first followed by pathway clustering and then visualization of DE evidence and c/d-scores plot).

1. **Pathway c-scores & d-scores calculation:**

   Three c-scores & d-scores parameters are required to be selected first.

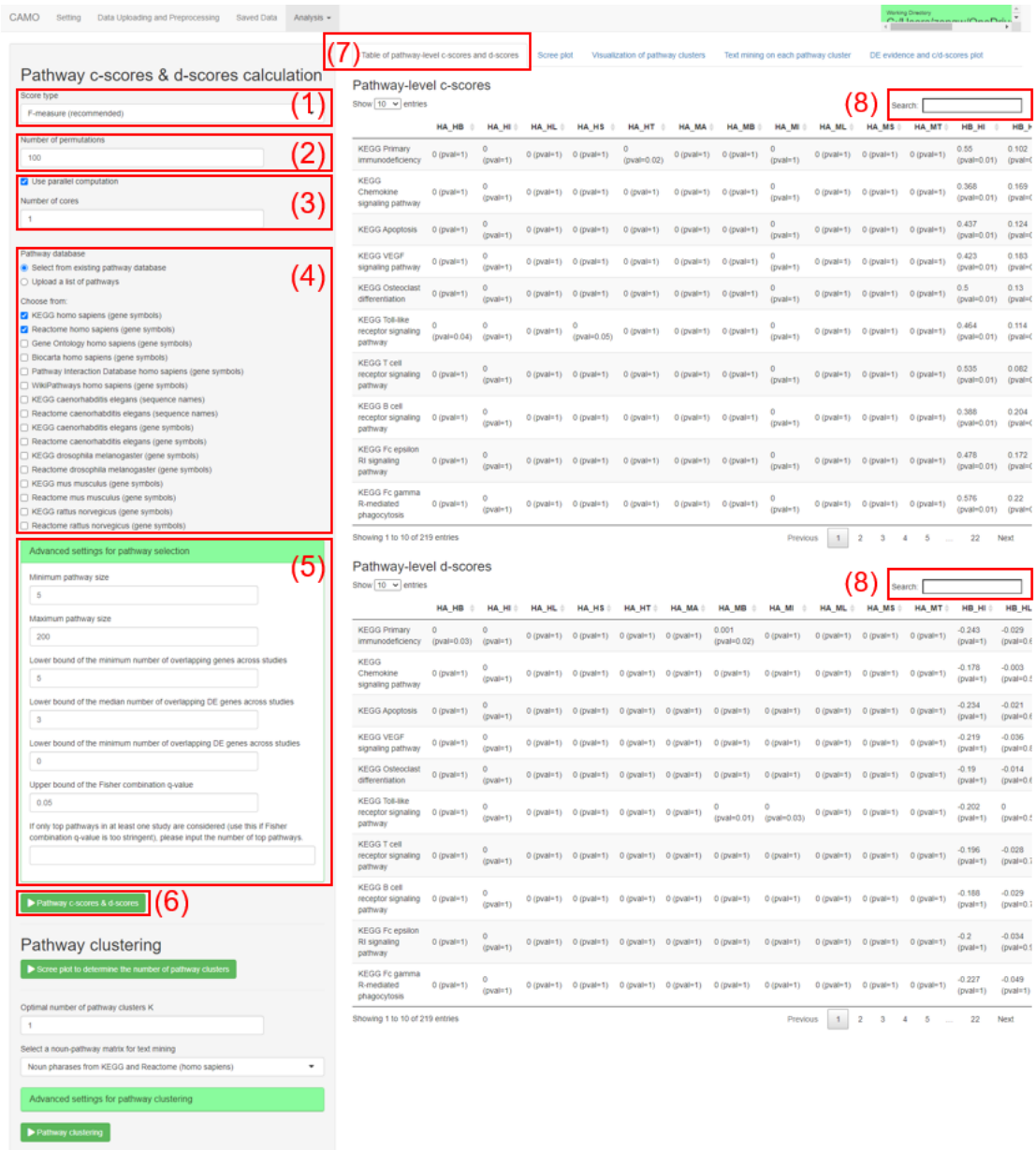

Figure 20: GUI for pathway level c-scores and d-scores calculation.

- "Score type" (see Figure 20 (1)). Three types of congruence scores defined in the CAMO paper are F-measure, Youden index and geometric mean. The definition of each score type is described in detail in the CAMO paper. F-measure is recommended because it is symmetric i.e., same value for a pair regardless of which is served as the reference group to be compared with.
- "Number of permutation" (see Figure 20 (2)) is the number of permutations during calculating the scores. 100 is used in our real data evaluation.
- "Use parallel computation" (see Figure 20 (3)). If the software is run on a server with multiple cores, parallel computing will speed up the process. CAMO will call R package *parallel* for parallel permutation.
- "Number of cores" (see Figure 20 (3)) is the number of cores used in the parallel computations of the "Use parallel computation" box is ticked.

Users need to select or upload pathway database of interests to perform pathway-level c-scores & d-scores computation on each of the pathway provided (see Figure 20 (4)). Please refer to the Section 3.2.2 for detailed pathway file preparation instruction. Among the pathways provided, an enrichment analysis will be run first to screen out pathways that are less relevant to the DE genes identified. By clicking "Advanced settings for pathway selection" (see Figure 20 (5)), users can adjust the pathway selection parameters in the collapse panel. Seven selection criteria are provided.

- "Minimum pathway size" and "Maximum pathway size". Only pathways whose size are within this range will be considered.
- "Lower bound of the minimum number of overlapping genes across studies". Only pathways who have the number of overlapping genes with each study to be greater than this value is considered.
- "Lower bound of median/minimum number of overlapping DE genes across studie". For each pathway, the number of genes appeared as DE genes for each study is calculated. Only pathways whose median/minimum across studies greater than this value is considered.
- "Upper bound of the Fisher combination q-value". Fisher meta-analysis is applied to the individual enrichment analysis to derive a meta-qvalue (Benjamini-Hochberg correction) for each pathway. Only pathways whose Fisher combination q-value are smaller than this value are considered.
- "If only top pathways in at least one study are considered (use this if meta-qvalue is too stringent), please input the number of top pathways.". Sometimes the applying a commonly used meta-qvalue threshold e.g.,0.05 can be too stringent because it considers enrichment across all studies. Alternatively, users can select the union of top enriched pathways in each individual study for analysis. This sets the top number of pathways in the individual enrichment analysis to be considered and a union of them will be selected.

If users are not sure about parameter selection, please leave them as default which generates good result in general based on our real data evaluation.

After setting the parameters described above, users can click on the 'Pathway c-scores & d-scores" button (Figure 20 (6)) to start the pathway c-scores & d-scores calculations. This step may take a while because of the permutation step. Users can check the process at the R console. Once the calculations is finished, summary tables of the c-scores and d-scores and their corresponding p-values for all study pairs are shown separately at the "Table of pathway-level c-scores d-scores" tab (see Figure 20 (7)) to the right. The closer the c/d-score is to 1, the more concordant/discordant the two studies are and vice versa. A c/d-score near to 0 indicates close to randomly generated values and less than 0 being even worse. Each row represents a pathway and their c/d-scores for each study pair (study names areconcatenated by "_" as column names) are shown in each corresponding cell with p-values in the brackets. Users can search for a particular pathway in the search boxes (see Figure 20 (8)). An aggregated c-scores & d-scores table for an individual pathway can also be generated in "Individual pathway analysis" (see Section 1).

The summary result table displayed in Figure 20 shows an example results generated from the human-mouse inflammatory models using default parameters for pathways from KEGG homo sapiens (gene

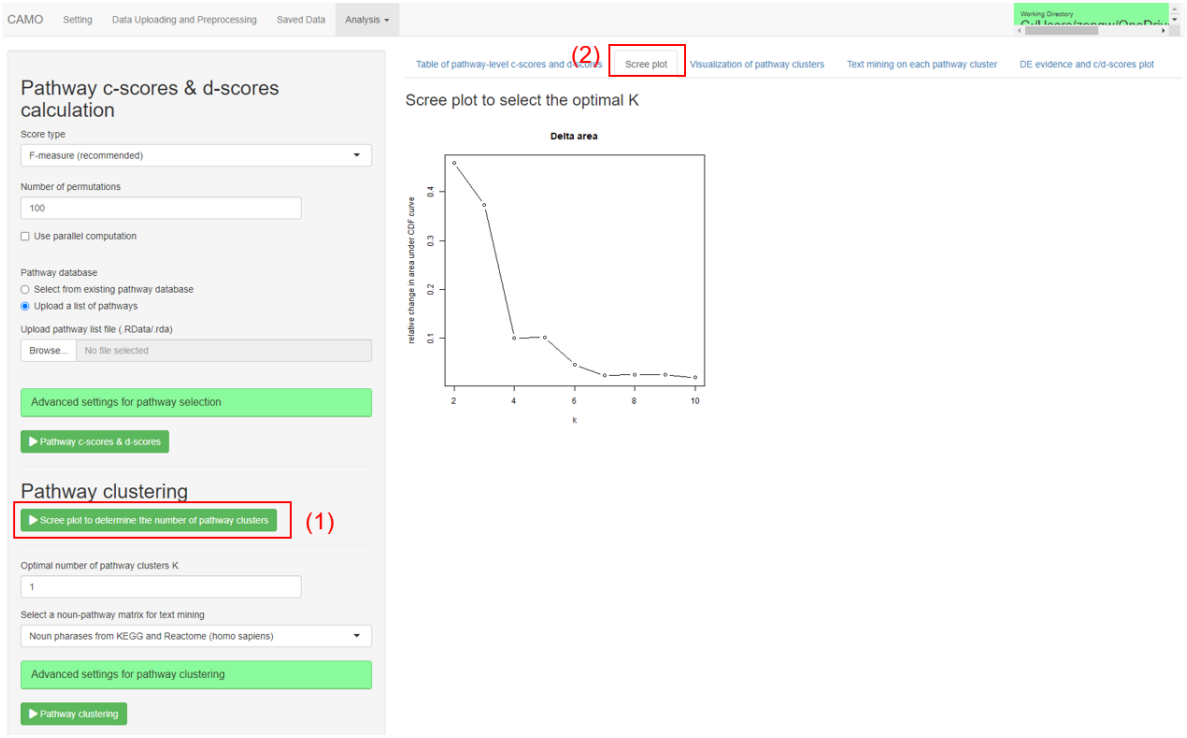

Figure 21: GUI for selecting the optimal cluster number.

symbols) and Reactome homo sapiens (gene symbols). Pathway-level c-socres & d-scores are saved as "ACS_ADS_pathway.RData" at the saving directory and a subdirectory "ACS_ADS_Pathway" is generated to store the c-socres & d-scores tables as .csv filesfor easy investigation.

2. **Pathway clustering:**

After the c-scores and d-scores been calculated, a consensus clustering with scatterness algorithm can be used to cluster pathways based on their c-scores. Here we only consider clustering using the concordance score because dissimilarity between multiple studies can hardly be defined using pairwise discordance scores due to the complexity of directions. To start with, users can click the "Scree plot to determine the number of pathway clusters" (see Figure 21 (1)) to generate a Scree plot at the "Scree plot" tab on the right (see Figure 21 (2)). As the number of clusters increases, the relative change in the area under the CDF curve tends to decrease. An optimal cluster number can be selected to be the k with small drop after moving to k+1. The cluster number selection is subjective as long as the results are justifiable, users can explore different k for meaningful results. The scree plot displayed in Figure 21 is generated from the human-mouse inflammatory models. Based on the this plot, we choose K=4 to be the optimal number of pathway clusters.

The selected number of clusters can be inputted at the "Optimal number of pathway clusters K" box (see Figure 22 (1)) to generate visualization of pathway clusters. To perform the text-mining, a noun-pathway matrix needs to be selected or prepared as instructed in Section 3.2.3 (see Figure 22 (2)). Additionally, two advanced parameters can be selected at the collapsed panel "Advanced settings for pathway clustering" (see Figure 22 (3)). The "Silhouette index cutoff to control scatterness" controls the scatterness level where larger values will result in tighter clusters. 0.1 usually generates good results. The "Lower bound of co-membership proportion shown in heatmap" control the color scheme in the co-membership heatmap at the "Text mining on each pathway cluster" tab.

By clicking on the "Pathway clustering" button (see Figure 22 (4)), CAMO will run pathway clustering

Figure 22: GUI for pathway clustering and text mining.

Figure 23: GUI for c/d-scores - DE evidence plot.

with scatterness and text mining. This step takes a few minutes to run. As a result, a MDS map of pathways colored by cluster labels and a heatmap of log-transformed pathway-level c-score p-values ordered by cluster labels will be shown in the "Visualization of pathway clusters" tab (see Figure 22 (5)). In the "Text mining on each pathway cluster" tab (see Figure 22 (6)), for each cluster, a table summarizing significant appeared noun phrases will be displayed with their enrichment q-values shown at the last column. Users can sort by the q-values from small to large to identify key phrases that are frequently appeared in the pathway descriptions in this cluster. This suggests the general topic covered by pathways in this cluster. In addition, the co-membership heatmap for each cluster is also shown in this tab so users can investigate which studies are more similar in terms of the topic suggested by the text mining results. Each cell of the co-membership heatmap is valued by the proportion pathways within which the two corresponding studies been clustered (clustering is based on the pairwise c-scores of this individual pathway) together among all pathways in this cluster. Large proportion indicates the the two corresponding studies are similar in a majority of pathways in this cluster. For better visualization, values below the "Cutting level for co-membership probability" (3) will be shadowed by blue in the co-membership heatmaps.

Figure 22 shows the results generated from the human-mouse inflammatory models. The clustering results (MDS map, heatmap and an aggregated text mining table) are saved at a subdirectory called "clustPathway" while the co-membership heatmaps are saved at the subdirectory "comemberPlot".

3. **Visualization of pathway DE evidence and c/d-scores:**

   This section visualizes DE evidence/strength from the Bayesian differential analysis and c/d scores at a pathway-level. Users can first select studies of interests for visualization at the bottom of the left side panel (see Figure 23 (1)) and then click the " DE evidence and c/d-scores" button (see Figure 23 (2)). A figure containing both the DE evidence and c/d-scores information for all study pairs generated the from the studies ticked at the "Select studies" will be displayed at the "DE evidence and c/d-scores" panel (see Figure 23 (3)) on the right main panel. In each subfigure, each dot is a pathway whose

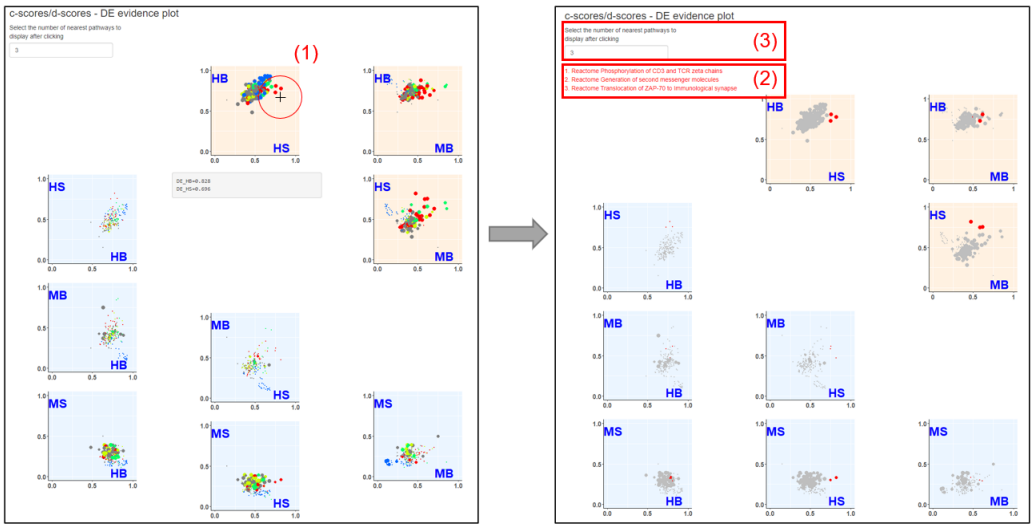

Figure 24: GUI for interactive c/d-scores - DE evidence plot exploration.

location is determined by the average posterior DE probability in the corresponding two studies (x-axis and y-axis) and size is determined by c-score (upper right orange figures) or d-score (lower left blue figures) p-values. Pathway dots are colored by their cluster labels. The posterior DE probabilities at a specific location will be shown at the bottom of the subfigure by hovering to the place. For further exploration, users can click on any locations and the plot will be updated such that the nearest pathways will colored by red (others are in grey) in all subfigures (see Figure 24 red circle (1)). Moreover, the names of the nearest pathways will be shown on top of the plot (see Figure 24 (2)). Users can change the number in the numerical input box (Figure 24 (3)) to control the number of nearest pathways to display after clicking. Note that users need to re-click on the plot to render an update figure after changing this number. The updating plot step may take a minute to run. This DE evidence and c/d-scores is aimed to assist users to identify individual pathways of interests for detailed exploration in the individual pathway analysis.

5.3 Individual pathway analysis

By clicking on the "Individual pathway analysis" under the "Analysis" tab, users are directed to the pathway-level congruence analysis page.

There are two main sections ("Individual pathway results browser", and "Save results for all pathways") on the left side panel. The "Individual pathway results browser" is used for investigating the individual pathway of interests whose results will be shown directly at the right main panels ("Table of pathway-level c-scores & d-scores", "Visualization of individual pathway", "Topological module detection for KEGG pathways - elbow plot", "Topological module detection for KEGG pathways - topological plot") while the "Save results for all pathways" will run all individual visualizations for all pathways and save results to the saving directory. The results will not be shown on the page but users can check R console for the process.

1. **Individual pathway results browser:**

   (a) Pathway-level c-scores & d-scores:

   By selecting a pathway of interests at the "Select a pathway:" box (see Figure 25 (1)), an aggregated c-scores & d-scores table of this pathway will be shown at the "Table of pathway-level c-scores & d-scores" tab. The upper triangular matrix (see Figure 25 (2)) contains the c-score (p-value) for the corresponding study pair (row vs column) of the selected pathway and the lower triangular matrix (see Figure 25 (3)) contains the d-score (p-value) for the corresponding study pair (row vs column) of the selected pathway. Users can switch to another pathway by choosing

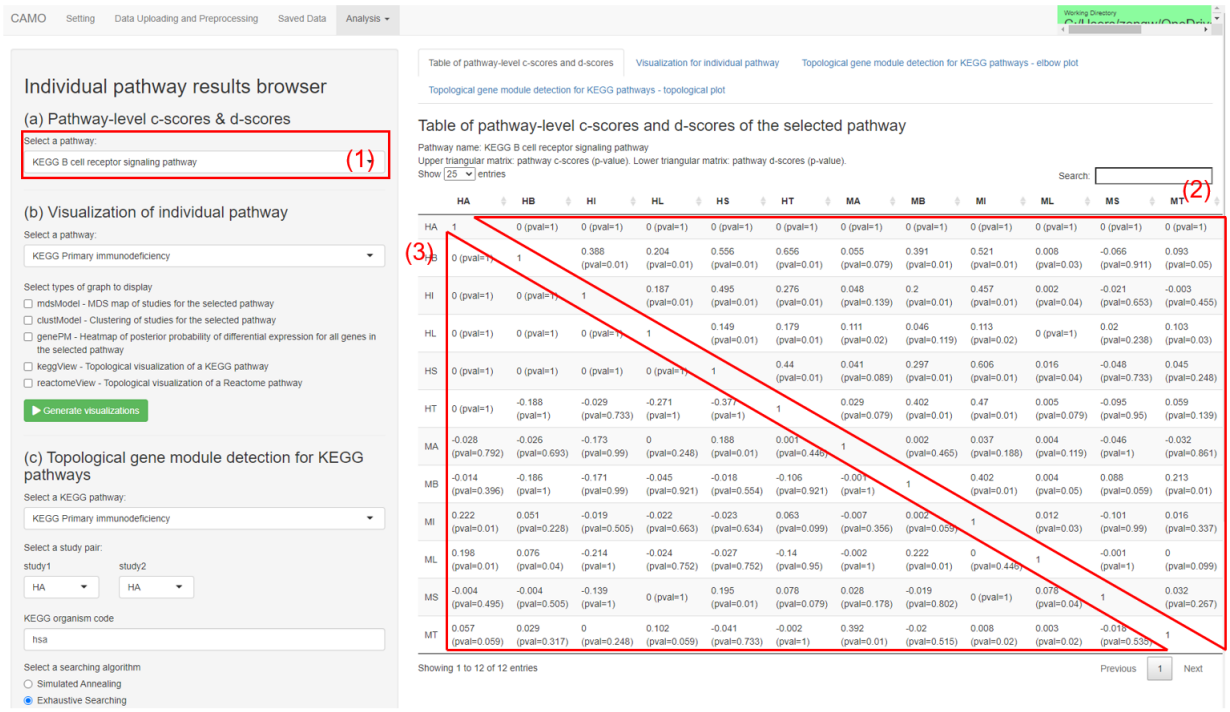

Figure 25: GUI for pathway level c-scores and d-scores of an individual pathway.

pathway from the drop-down menu or typing the pathway name in Figure 25 (1) and the table will be updated immediately.

(b) Visualization of an individual pathway:

Various visualization tools are provided at individual pathway level to facilitate pathway knowledge retrieval (see Figure 26 (2)), including

- "mdsModel": MDS map of all studies based on the c-scores of the selected pathway. This demonstrates the distance between studies suggested by genes this pathway.
- "clustModel": Heatmap of the c-scores in the selected pathway. Studies (rows and columns) are ordered based on clustering using pairwise c-scores.
- "genePM": Heatmap of posterior DE probability all genes in the selected pathway.
- "keggView": Topological visualization of a KEGG pathway for each pair of studies. Nodes are molecular objects (genes, proteins, small molecules, etc.) and edges are molecular interactions/reactions/relations. Nodes containing genes appeared in data are bi-colored based on the average posterior DE probability in each of the study in the study pair to compare the DE strength/evidence between the two studies. If this is selected, "Settings for topological visualization" will be shown as a collapsed panel (see Figure 26 (3)). Users need to specify the KEGG organism code of the reference species which can be find at `https://www.genome.jp/kegg/catalog/org_list.html`. Since CAMO calls the R package *pathview* for KEGG topological visualization which requires genes to be named by their entrez IDs. If data genes use a different type of names, the "Gene names in data matrix are Entrez IDs?" should be selected to be FALSE and users can either upload their own matching file or allow CAMO to automatically retrieve entrez IDs by Bioconductor packages (see Section3.2.4 for detailed instruction). KEGG topological plots will be generated for all study pairs constructed from studies selected at the "Select a subset of studies to generalize KEGG/Reactome topology plots".
- "reactomeView": Topological visualization of a Reactome pathway for each pair of studies. Nodes are molecular objects (genes, proteins, small molecules, etc.) and edges are molecular interactions/reactions/relations. Nodes containing genes appeared in data are bi-colored based on the average posterior DE probability in each of the study in the study pair to compare the

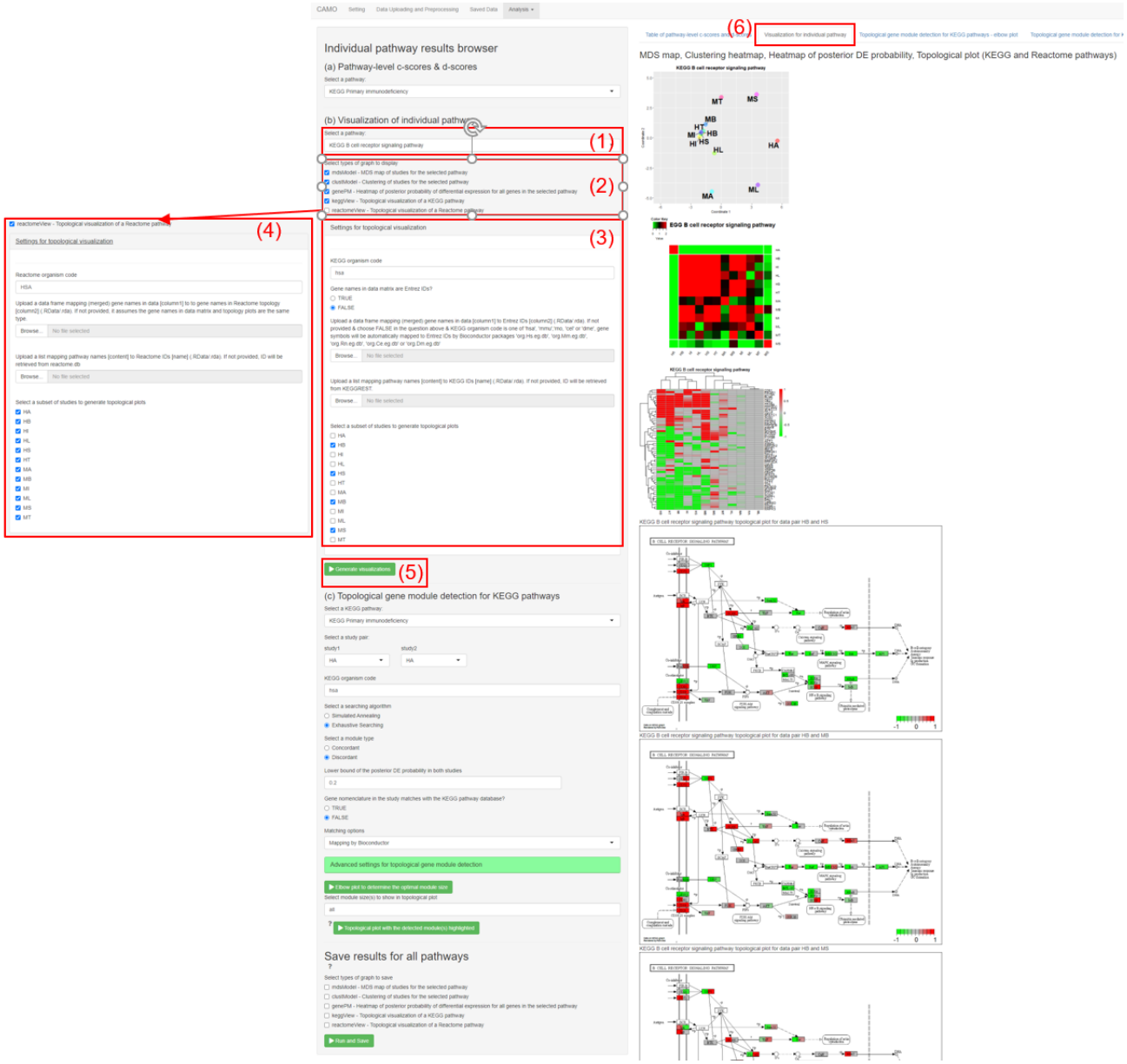

Figure 26: GUI for individual pathway visualization.

DE strength/evidence between the two studies. If this is selected, "Settings for topological visualization" will be shown as a collapsed panel (see Figure 26 (4)). Users need to specify the Reactome organism code of the reference species which can be find by first clicking on the corresponding identifier at `https://reactome.org/content/schema/objects/Species` and the Reactome organism code is the three-character abbreviation on the species page. Since CAMO assumes that the data gene names are consistent with gene names on the Reactome topological figures downloaded from the Reactome website which are usually gene symbols. If this is not the case, users can either upload their own matching file following the instruction in Section3.2.4. Reactome topological plots will be generated for all study pairs constructed from studies selected at the "Select a subset of studies to generalize KEGG/Reactome topology plots".

Users can select a pathway of interests at (see Figure 26 (1)) and click on the "Generate visualizations" (see Figure 26 (5) to initiate the analysis. The resulted figures will be shown at the "Visualization of individual pathway" on the right main panel (see Figure 26 (6)) and also saved to the saving directory.

(c) Topological gene module detection for KEGG pathways:

If a KEGG pathway has a large number of concordant/discordant nodes with strong DE evidence, users can further investigate their structure by running the topological gene module detection algorithm to identify locally related/interactive genes within a pathway reaction map. Several parameters/input are required for the topological gene module detection (see Figure 27 (1)), including

- "Select a KEGG pathway": Select a KEGG pathway of interests for module detection.
- "Select a study pair": Since the module detection is to identify local modules among all pairwise concordant/discordant nodes, users need to specify a study pair to define the concordant/discordant nodes. Users can use the "keggView" tool to identify study pairs with large number of concordant/discordant nodes in the Visualization of an individual pathway section before module detection.
- "KEGG organism code": Specify the KEGG organism code of the reference species which can be find at `https://www.genome.jp/kegg/catalog/org_list.html`.
- "Select a searching algorithm": Two searching schemes are provided to identify gene modules with the smallest average shortest path at a given module size. Exhaustive search algorithm is feasible for a pathway with a limited number of concordant/discordant genes (e.g., size¡=30). For a pathway with a number of concordant/discordant genes greater than 30, the simulated annealing (SA) algorithm is suggested for fast search.
- "Select a module type": Whether the gene module detection is within the concordant genes or discordant genes.
- "Lower bound of the posterior DE probability in both studies": Concordant/discordant genes (searching space) are defined as genes whose posterior DE mean having absolute value greater than this value and the signs are in the same/opposite directions in the study pair selected. This is set because the concordance/discordance definition is only meaningful when the DE signal is greater than a certain threshold in both studies. 0.2 works well in general but users can increase/reduce to expand/limit the searching space.
- "Gene nomenclature in the study matches with the KEGG pathway database?" and "Matching options": If the entry names used in the KEGG topological plot (.xml file on KEGG website) is different from the data gene names, the first questioin should be FALSE. The entry names can be checked by clicking nodes on the interactive topological figures on the KEGG website. Users need to either upload their own matching file or allow CAMO to automatically matching them following the instruction in Section 3.2.4.

In addition, users can tweak the algorithm by adjusting the parameters in the collapsed penal "Advanced settings for topological gene module detection" (see Figure 27 (2)). The default parameters in the panel work well under real data evaluation and users should only change them if they understand the algorithm well and have a specific purpose. Here are parameter definitions:

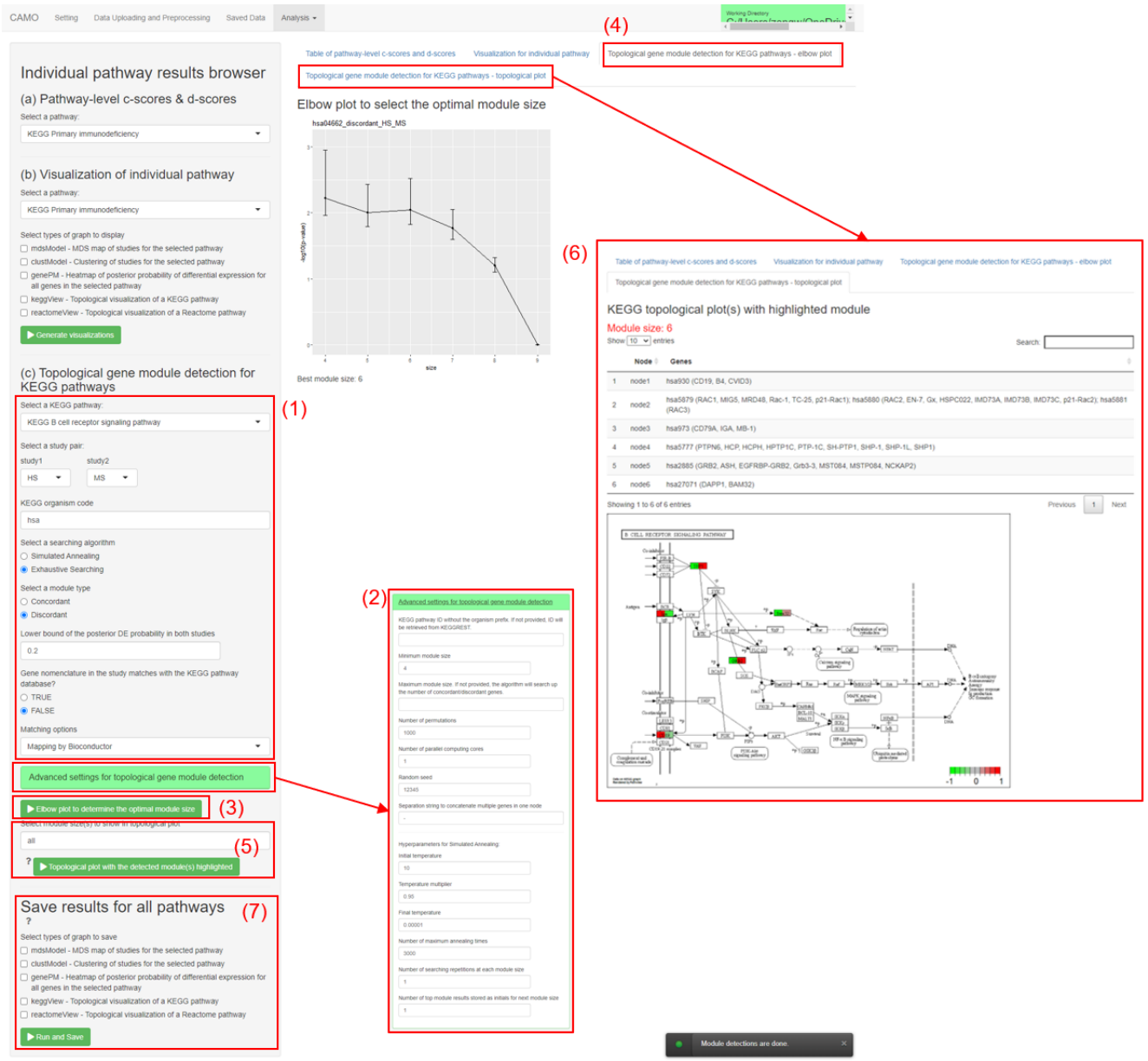

Figure 27: GUI for topological gene module detection and saving results for all pathways.

- "KEGG pathway ID without the organism prefix.": KEGG pathway ID is a required input for module detection algorithm which can be retrieved by R package *KEGGREST*. However, sometimes the pathway name may change resulted in failure to find the corresponding pathway ID. If this happened, users are encourage to manually input the corresponding KEGG pathway ID for the pathway selected at "Select a KEGG pathway". Please note that the pathway ID should not include the organism prefix e.g., "04662" instead of "hsa04662" for pathway "B cell receptor signaling pathway - Homo sapiens (human)".
- "Minimum module size" and "Maximum module size": The algorithm will search modules whose sizes are within this range. If the maximum module size is not provided, it will automatically search up to the number of all concordant/discordant genes satisfying the DE probability critrion in this study pair.
- "Number of permutations", "Number of parallel computing cores" and "Random seed": Permutation parameters to generate p-values of the observed smallest average shortest path at each module size.
- "Separation string to concatenate multiple genes in one node": In the internal process, when a single node represents multiple genes, genes will appeared as a single row and a single column in the shortest path matrix and named after gene names concatenated by this symbol. "-" is used as a default to avoid confusion with "_" contained in some gene names e.g., cel gene names start with "CELE_".
- "Initial temperature", "Temperature multiplier", "Final temperature" and "Number of maximum annealing times": These are basic simulated annealing parameters. Intuitively, the initial temperature controls the acceptance of a trial assignment, the number of maximum annealing times controls the maximum number of annealing iterations, the temperature multiplier controls speed of cooling down process for the initial temperature to drop below the final temperature and stop the process. The annealing process is harder when the initial temperature is larger, the temperature multiplier is smaller and the number of maximum annealing times is larger.
- "Number of searching repetitions at each module size" and "Number of top module results stored as initials for next module size": When applying the SA algorithm to a spectrum of module sizes, to further improve the performance, at each module size m, the algorithm runs x times and the top y results are stored and passed to the next m+1 scenario as initials. Borrowing initial values from the previous step allows this procedure to converge faster and y>1 helps to robustize the procedure when multiple close-to-optimal solutions exist. Here the number of searching repetitions at each module size is x and the number of top module results stored as initials for next module size is y.

By clicking on the "Elbow plot to determine the optimal module size" (see Figure 27 (3)), the gene module detection algorithm will be run for a spectrum of module size from the "Minimum module size" to the "Maximum module size" specified above and an elbow plot of the -log10 transformed permuted p-values of the observed average smallest shorted path will be shown at the "Topological module detection for KEGG pathways - elbow plot" tab (see Figure 27 (4)) on the right main panel. The confidence interval of the p-values with two standard deviation is shown as a vertical bar at each module size. An optimal module size is suggested below the plot as the maximum module size whose p-value is within 2 standard deviations of the minimum p-value. However, users should consider the p-value elbow plot, KEGG topology plots together with their biological insights in determining an appropriate module size.

Users then can select the module sizes of interests by typing in numbers separated by "," e.g.,"1,2,3" at (5) to generate KEGG topological plots with highlighted modules at the module sizes selected. Typing "all" if all module sizes are of interests. By clicking on the "Topological plots with the detected module(s) highlighted", the topological plots will be shown at the "Topological module detection for KEGG pathways - topological plot" on the right panel (see Figure 27 (6)).

2. **Save results for all pathways:**

This section is to generate all visualization results for all pathways and save them to local directory. The tools in this section is named same as in the (b) Visualization of an individual pathway in the Individual pathway results browser. By clicking on the "Save and Run" (see Figure 27 (7)), the results

will be automatically saved to the saving directory and not show up on the main panel. Users can check the running process at the R console.

6 Complete list of options

6.1 Data Uploading and Preprocessing

Complete List of Options:

1. Upload data:
   - Select a species: homo sapiens (hs)/mus musculus (mm)/rattus norvegicus (rn)/caenorhabditis elegans (ce)/drosophila melanogaster (dm)/other. Select the species of the data to upload.
   - Specify the species name: Type in the user-defined species name.
   - Input data type:
     (a) Precalculated p-value and log fold change (logFC) from gene-level differential expression analysis:
       – Upload p-value and logFC file (.csv): Uploda the .csv file including the p-values and logFC values from DE analysis of the study data from the local directory.
     (b) Preprocessed gene expression data:
       – Upload gene expression data file (.csv): Upload the gene-expression data file from the local directory
       – Upload clinical file (.csv): Upload the clinical file including the case or control information from the local directory.
       – Case name: Select which group label in clinical file represents cases.
       – Differential analysis method: LIMMA/DEseq2. Select the type of differential analysis to derive the p-values and logFC.
   - Bayesian differential analysis: Calculate the posterior DE prbability using a threshold-free Bayesian approach.
2. Save single study:
   - Study name (Please do not use '_' in study name):: Type in the name of the uploaded data, this will be shown as the StudyName column at the "Saved Data" page.
   - Save: Save the Bayesian differential analysis results into the study database.

6.2 Saved Data

Complete List of Options:

- Selected studies: Click on the rows of the "List of saved data" to select studies for merging/deleting.
- Ortholog file: Same species - skip orholog matching/homo sapiens (hs) vs mus musculus (mm)/homo sapiens (hs) vs rattus norvegicus (rn)/homo sapiens (hs) vs caenorhabditis elegans (ce)/homo sapiens (hs) vs drosophila melanogaster (dm)/caenorhabditis elegans (ce) vs drosophila melanogaster (dm)/Upload an orthologs file. Select the ortholog file used for merging cross-species studies. The ortholog file can be previewed at the "Orthologs file selected" on the right.
- Upload an ortholog file (.RData/.rda): Upload an orthologou file from local directory.
- Reference species: Select the species used as a reference during merging.
- Match and merge: Match and merge all selected studies.
- Delete selected studies: Delete the selected studies permanently.

6.3 Analysis: Genome-wide analysis

Complete List of Options:

- Score type: F-measure/Youden index/Geometric mean. Select the measure type used for deriving the c-scores and d-scores. F-measure as the only symmetric measure is recommended.
- Number of permutations: Type in the number of permutations used to deriving p-values of c-scores and d-scores.
- Genome-wide c-scores & d-scores: Start genome-wide c-scores and d-scores computation.
- Genome-wide MDS map: Generates the MDS plot using genome-wide c-scores.

6.4 Analysis: Pathway-based analysis

Complete List of Options:

1. Pathway c-scores & d-scores calculation:
    - Score type: F-measure/Youden index/Geometric mean. Select the measure type used for deriving the c-scores and d-scores. F-measure as the only symmetric measure is recommended.
    - Number of permutations: Type in the number of permutations used to deriving p-values of c-scores and d-scores.
    - Use parallel computation: Select if parallel computation can be used for fast computing.
    - Number of cores: Type the number of cores used for parallel computing.
    - Pathway database:
        - Select from existing pathway database: 16 pathway databases for 5 species are provided, i.e., KEGG homo sapiens (gene symbols), Reactome homo sapiens (gene symbols), Gene Ontology homo sapiens (gene symbols), Biocarta homo sapiens (gene symbols), Pathway Interaction Database homo sapiens (gene symbols), WikiPathways homo sapiens (gene symbols), KEGG caenorhabditis elegans (sequence names), Reactome caenorhabditis elegans (sequence names), KEGG caenorhabditis elegans (gene symbols), Reactome caenorhabditis elegans (gene symbols), KEGG drosophila melanogaster (gene symbols), Reactome drosophila melanogaster (gene symbols), KEGG mus musculus (gene symbols), Reactome mus musculus (gene symbols), KEGG rattus norvegicus (gene symbols), Reactome rattus norvegicus (gene symbols). Select pathways for pathway-level c-scores and d-scores calculation. Multiple selection is allowed.
        - Upload a list of pathways: Upload a ".RData/.rda" containing a list of gene sets from local directory.
    - Advanced settings for pathway selection:
        - Minimum pathway size.
        - Maximum pathway size.
        - Lower bound of the minimum number of overlapping genes across studies.
        - Lower bound of the median number of overlapping DE genes across studies.
        - Lower bound of the minimum number of overlapping DE genes across studies.
        - Upper bound of the Fisher meta-qvalue.
        - The number of top pathways requirement in individual studies.
    - Pathway c-scores & d-scores: Start genome-wide c-scores and d-scores computation.
2. Pathway clustering
    - Scree plot to determine the number of clusters: Generate a scree plot of the relative change in the area under the CDF curve over the number of clusters to determine the optimal number of clusters.

- Optimal number of pathway clusters K: Type in a number as the number of clusters desired to have.
- Select a noun-pathway matrix for text mining: Noun pharases from KEGG and Reactome (homo sapiens/mus musculus/rattus norvegicus/caenorhabditis elegansdrosophila melanogaster)/Upload a noun-pathway matrix file/Skip text mining. Select from prepared matrices, uploading own matrix or skip text mining in pathway clustering.
- Upload a noun-pathway file for text mining (.RData/.rda): Upload a user-defined flat noun-pathway matrix.
- Advanced settings for pathway clustering:
  - Silhouette index cutoff to control scatterness: Type in a Silhouette lower bound to control scatterness level. Results will have more scattered pathways given a larger value.
  - Lower bound of co-membership proportion shown in heatmap: Type in a proportion value and any value below this will be colored blue in the co-membership heatmap.
  - Pathway clustering: Start clustering, text mining and calculating co-membership probability.

3. Visualization of pathway DE evidence and c/d-scores:

- Select studies: options are all merges studies. Select studies for DE evidence and c/d-scores plot.
- DE evidence and c/d-scores plot: Start generating DE evidence and c/d-scores plot plot.

6.5 Analysis: Individual pathway analysis

Complete List of Options:

1. Individual pathway results browser:

   (a) Pathway-level c-scores & d-scores:
   - Select a pathway: Select a pathway from the scroll-down list to show its c-scores/d-scores. Users can type in keywords and select from suggested pathways.

   (b) Visualization of individual pathway:
   - Select a pathway: Select a pathway from the scroll-down list. Users can type in keywords and select from suggested pathways.
   - Select types of graph to display: mdsModel/clustModel/genePM/keggView/reactomeView. Select tools for visualization. Multiple selections are allowed.
   - Settings for topological visualization:
     - KEGG organism code
     - Gene names in data matrix are Entrez IDs?
     - Upload a data frame mapping (merged) gene names in data [column1] to Entrez IDs [column2] (.RData/.rda)
     - Upload a list mapping pathway names [content] to KEGG IDs [name] (.RData/.rda).
     - Reactome organism code
     - Upload a data frame mapping (merged) gene names in data [column1] to to gene names in Reactome topology [column2] (.RData/.rda)
     - Upload a list mapping pathway names [content] to Reactome IDs [name] (.RData/.rda)
     - Select a subset of studies to generate topological plots: Topological plots will be generated for pairwise studies constructed from the selected ones.
   - Generate visualizations: Start to run the tools selected.

   (c) Topological gene module detection for KEGG pathways
   - Select a KEGG pathway: Select a pathway from the scroll-down list. Users can type in keywords and select from suggested pathways.
   - Select a study pair: Select two different studies.
   - Select a searching algorithm: Simulated Annealing/Exhaustive Searching.

- Select a module type: Concordant/Discordant.
- Lower bound of the posterior DE probability in both studies
- Gene nomenclature in the study matches with the KEGG pathway database?
- Matching options: Mapping by Bioconductor/Upload a mapping file.
- Upload a data frame to match gene names in studies to KEGG topology plots (.RData/.rda)
- Advanced settings for topological gene module detection:
  - KEGG pathway ID without the organism prefix
  - Minimum module size
  - Maximum module size
  - Number of permutations
  - Number of parallel computing cores
  - Random seed
  - Separation string to concatenate multiple genes in one node
  - Initial temperature
  - Temperature multiplier
  - Final temperature
  - Number of maximum annealing times
  - Number of searching repetitions at each module size
  - Number of top module results stored as initials for next module size
- Elbow plot to determine the optimal module size: Start gene module detection at each selected module size.
- Select module size(s) to show in topological plot: Type in module sizes of interests to visualize their location on the topological plots. Numbers should be seperated by "," without space in between.
- Topological plot with the detected module(s) highlighted: Start generating topological plots.

2. Save results for all pathways:

   - Select types of graph to save: mdsModel/clustModel/genePM/keggView/reactomeView. Select tools to run and save for all pathways. Multiple selections are allowed.
   - Settings for topological visualization:
     - KEGG organism code
     - Gene names in data matrix are Entrez IDs?
     - Upload a data frame mapping (merged) gene names in data [column1] to Entrez IDs [column2] (.RData/.rda)
     - Upload a list mapping pathway names [content] to KEGG IDs [name] (.RData/.rda).
     - Reactome organism code
     - Upload a data frame mapping (merged) gene names in data [column1] to to gene names in Reactome topology [column2] (.RData/.rda)
     - Upload a list mapping pathway names [content] to Reactome IDs [name] (.RData/.rda)
     - Select a subset of studies to generate topological plots: Topological plots will be generated for pairwise studies constructed from the selected ones.
   - Run and Save: Start to run and save the tools selected.

References

Li, J. J., Huang, H., Bickel, P. J., and Brenner, S. E. (2014). Comparison of d. melanogaster and c. elegans developmental stages, tissues, and cells by modencode rna-seq data. *Genome research*, 24(7):1086–1101.

Zeng, X., Zong, W., Lin, C.-W., Fang, Z., Ma, T., Lewis, D. A., Enwright, J. F., and Tseng, G. C. (2020). Comparative pathway integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis. *Genes*, 11(6):696.

Zong, W., Rahman, M. T., Zhu, L., Zeng, X., Zhang, Y., Zou, J., Liu, S., Ren, Z., Li, J. J., Oesterreich, S., et al. (2021). Camo: A molecular congruence analysis framework for evaluating model organisms. *bioRxiv*.