— title: “MutRank” author: “Elly Poretsky, Alisa Huffaker” date: “April 13, 2020” output: html_document: toc: true # table of content true toc_depth: 3 # upto three depths of headings (specified by #, ## and ###) number_sections: true ## if you want number sections at each table header theme: united # many options for theme, this one is my favorite. highlight: tango # specifies the syntax highlighting style —

Contents

1. Introduction

# # #

With reduced cost and increased accessibility of next generation sequencing technologies, public and private custom large scale transcriptomic datasets are now commonly analyzed by many laboratories. For example, in plants many studies and online databases combine numerous transcriptomic samples from species, genotypes, developmental stages, tissues and physiological conditions to understand traits of agronomic significance. The publication of transcriptomes from thousands of plant species are expected to speed large-scale transcriptomic experiments in non-model organisms. Transcriptomic data can help unravel complex biological processes in part through understanding gene coexpression analyses. Often genes that function together within a pathway have a higher probability of being transcriptionally coregulated and can help make early predictions of meaningful associations and gene functions.

Many databases and webtools have been developed to facilitate coexpression analyses and often use Pearson's Correlation Coefficient (PCC) as a measure of coexpression. Mutual Rank (MR), the geometric mean of the ranked PCCs between a pair of genes, has been proposed as an alternative measure of coexpression to PCC. MR was shown to be better at predicting gene function compared to PCC independent of how the PCC coexpression database was constructed and of the reference gene tested. When the MR- and PCC-based coexpression databases of multiple plant species were converted into coexpression networks, the MR-based coexpression networks were more comparable than PCC-based coexpression networks across species using different metrics. Clustering of the MR-based networks produced clusters that were enriched for enzymes associated with plant specialized metabolism pathways. Confirmed through diverse empirical approaches, targeted MR-based coexpression analyses were recently leveraged as powerful tools enabling the narrowing of candidates and accurate prediction of genes encoding enzymes in maize specialized metabolism including the kauralexin and zealexin pathways.

Despite the usefulness of existing coexpression databases few databases enable flexible hypothesis testing and tool-based simplicity integrating user-provided expression data and supporting information. Integrating user-provided supporting information with coexpression results can facilitate the prediction of meaningful functional associations and tentative assignment of putative gene functions. We developed a R Shiny web-application, termed MutRank, to facilitate exploratory targeted MR-based coexpression analyses. Using the R Shiny framework allowed for the design of a coexpression analysis platform that utilizes useful R packages in addition to incorporating user-provided expression data and supporting information. A web-application is also advantageous for generating a highly customizable and easy-to-use interface that can run on most personal computers. In addition to identifying the most highly coexpressed genes in any user-provided expression dataset, MutRank integrates supporting information such as gene annotations, differential-expression data, predicted domains and assigned GO terms and provides useful tabular and graphical outputs as foundation for empirical hypothesis testing.

2. Getting Started

2.1. Requirements

- `R` - https://cran.r-project.org/src/base/R-3/ * `R Stduio` - https://rstudio.com/products/rstudio/download/ * `Java` (requires restarting) - https://java.com/en/download/

2.2. Installation

1. Download or clone MutRank from: https://github.com/eporetsky/mutRank 2. Unzip and open the `app.R` file using R Studio 3. To start MutRank press the `Run App` button in R Studio 4. Start using MutRank in the browser or window mode * When MutRank first starts it installs and loads required R libraries

2.3. R Dependencies

MutRank will automatically install these packages when you start it for the first time.

- hypergea_1.3.6 * ontologyIndex_2.5 * reshape2_1.4.3 * RColorBrewer_1.1-2 * data.table_1.12.8 * ggplot2_3.3.0 * visNetwork_2.0.9 *igraph_1.2.4.2* shinythemes_1.1.2 * shiny_1.4.0.2

2.4. File Formats

1. Comma-separated values (.csv): Expression and differential expression data 2. Tab-separated values (.tsv) - Annotations, symbols, Pfam domains, GO assignments and custom cetegories

2.5. Data Preparation

In the main MutRank application folder exist 7 separate folders that contain the example expression data file and the supporting information files. Use the formatting in the provided example files when preparing custom files, preferably by viewing the files using a simple text editor. It is important to correctly format the file and use filename extensions (i.e. csv and tsv) so they can be included in the selection menus when MutRank starts and used for the coexpression analyses.

3. Navigating the MutRank Tabs

3.1. `Data Input` Tab

The Data Input tab is the first tab of MutRank in which users can load their expression data and supporting information for MR-based coexpression analyses (Fig. 1). MutRank only requires expression data to calculate MR values between genes and will integrate coexpression results with user-provided supporting information. When starting MutRank it automatically lists all the files in the different folders that end with the correct filename extensions and adds them to the dropdown menus. Alternatively, we included an option to temporarily upload files manually through a file browser, useful if MutRunk is hosted on a server using shiny-server. MutRank will load the last selected file or uploaded file. When loading large expression data files (and the large go-basic.obo Gene Onthology database file) you might experience a short delay. We have added a short text output under the expression data loading section that will update once the expression data file is loaded and state the number of rows and columns in the data. By default MutRank starts with loading the example files. This can be changed by pressing the Save Default button after selecting your preferred default files or manually editing the default_files.csv file.

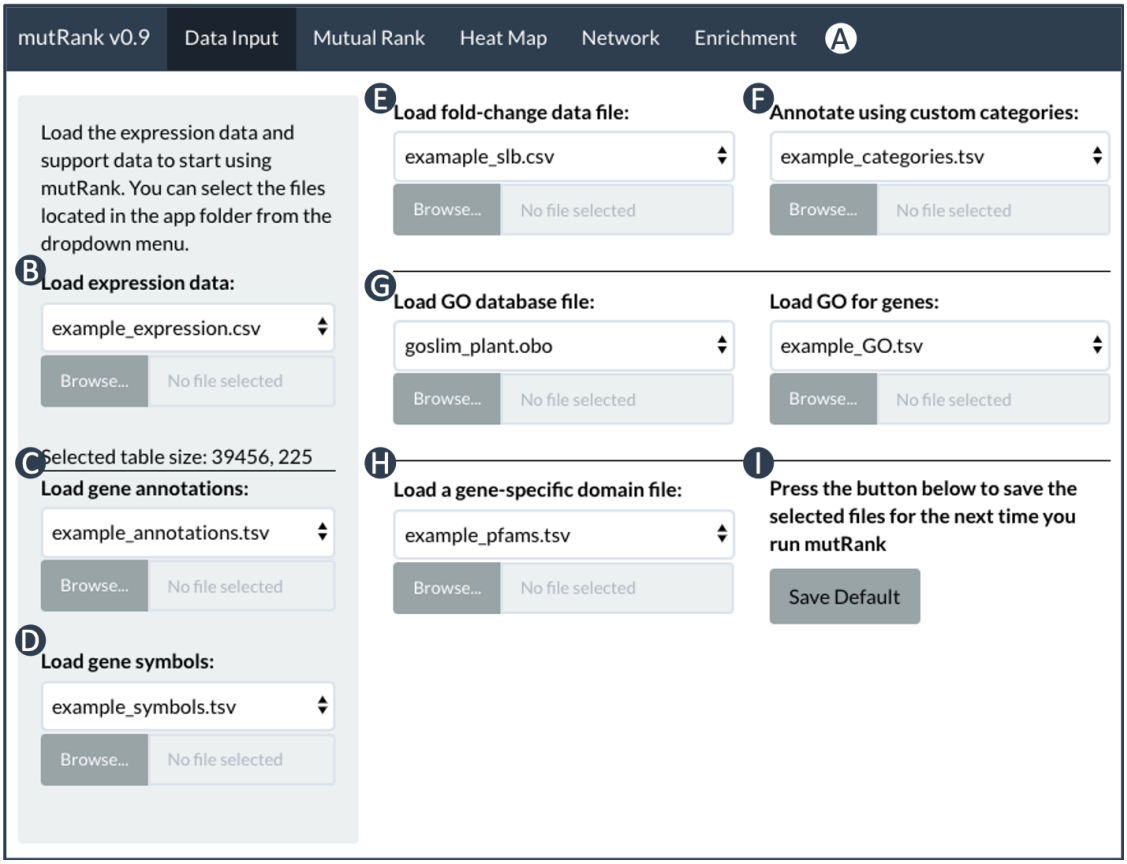


Figure 1: Overview of MutRank 1.0, Data Input interface and the selection of custom user defined datasets. (A) Cursor based navigation between the different component pages of MutRank enabled by the top tab panel. In the left side panel of the Data Input page you can load (B) expression data, (C) gene annotations and (D) gene symbols. In the main panel you can load (E) differential expression data, (F) custom categories, (G) GO database and GO assignments, (H) protein domain assignments and (I) to save default files to open upon the next MutRank analyses run.

3.2. `Mutual Rank` Tab

Once the expression data and supporting information are loaded you can start the coexpression analyses using a reference gene or gene list (Fig. 2). First select from one of three methods for the reference gene(s): (1) Single reference gene, (2) compound reference gene and (3) reference gene list. The compound reference gene method creates a compound reference gene from the average, sum, maximum or minimum expression values of the reference gene list and the reference gene list method calculates the MR values between the genes in the list. Genes in both gene lists can be separated by: tab, new line, vertical tab, space and comma. In the example below we selected the maize reference gene termed benzoxazinoidless 1 (bx1; GRMZM2G085381) which encodes an Indole-3-glycerol phosphate lyase (Fig. 2). MutRank only calculates MR values for a set number of the top coexpressed genes based on PCC values. This practical trade-off between whole genome and targeted coexpression analyses allows MutRank to rapidly complete the analyses and to run on the resources of most personal computers.

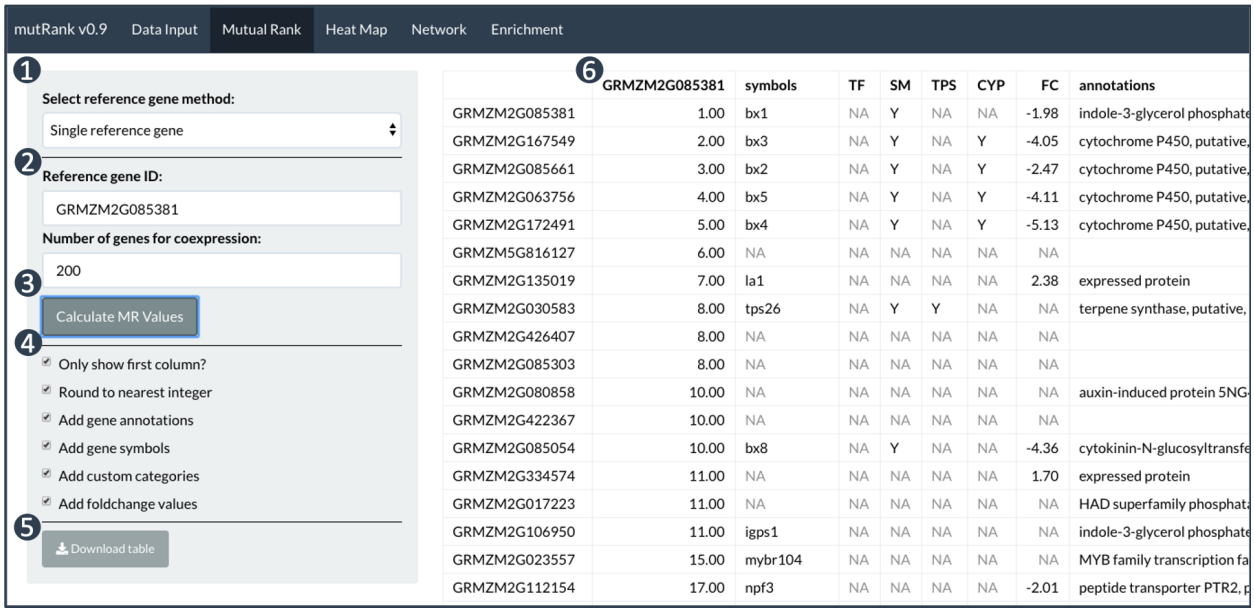


Figure 2: Mutual Rank 1.0 calculation interface and the selection of user defined references genes, families and lists. (A) The user starts by selecting one of the three reference-gene methods. (B) Reference gene(s) are inserted into the box and (C) the desired number of the top coexpressed genes (based on PCC) to include is defined followed by a button "Calculate MR Values" to start calculating the MR coexpression table. (D) A series of additional output options can be further selected to provide supporting information integrated with the coexpression table output. (E) Users can then download the custom coexpression table (F) as it is displayed.

3.3. Coexpression `Heatmap` Tab

The MR-based coexpression table results generated in the previous tab can be used to generate different graphical outputs in the Heat Map tab (Fig. 3). Often it is desirable to rapidly examine results using a coexpression heatmap in which a color gradient is used to represent the level of coexpression and only display MR values below a select threshold. For practical and useful visual results, we recommend using a small/limited number of genes for generating the coexpression heatmap.

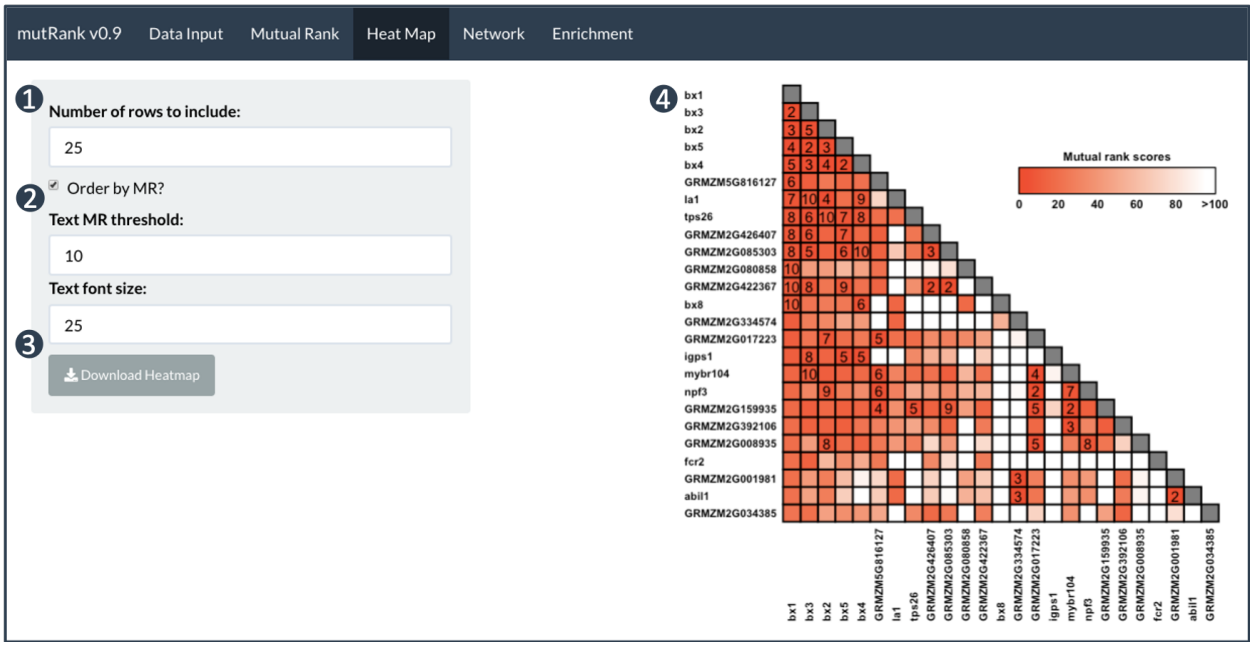


Figure 3: Overview of the custom Heat Map interface of MutRank 1.0 demonstrates simple and useful control over graphical outputs. (A) Selected number of comparatively lowest MR score genes to include after the reference gene (limited to 25). Note: the Mutual Rank tab enables much larger MR data outputs. (B) Users can select the display threshold and font size for the appearance of MR values within the custom heatmap. (C) Download the coexpression heatmap as a .png file.

3.4. Coexpression `Network` Tab

The coexpression network uses the MR-based coexpression table as an adjacency matrix input to create an igraph network instance (Fig. 4). The nodes of the network are the genes in the coexpression table and edges connect nodes MR value between to genes is lower than the MR threshold (this needs to be clear, I get lost every time I read it. . .). By default the reference gene is assigned star icon and all other genes are circle shaped. Some of the supporting information is set as node attributes by igraph. The igraph network is converted to a dynamic java-script netowrk visualization using the vizNetwork package. Gene IDs are written under the nodes and when present symbols automatically replace the full gene IDs. Gene annotations can be dynamically viewed by selecting the gene node. Select a column from your differential expression data to change the color of the gene nodes. MutRank currently supports log2 fold-change values and plots them as 3 grades of blue colors for down-regulated genes and 3 grades of red colors for up-regulated genes. For user-defined categories select a column from the categoreis for one of the 5 shapes. If a gene belongs to more than one category, the shapes lower in the list will overwrite the previous shapes. Use the options in the side panel to edit the network visualizations. (This section still needs some help!)

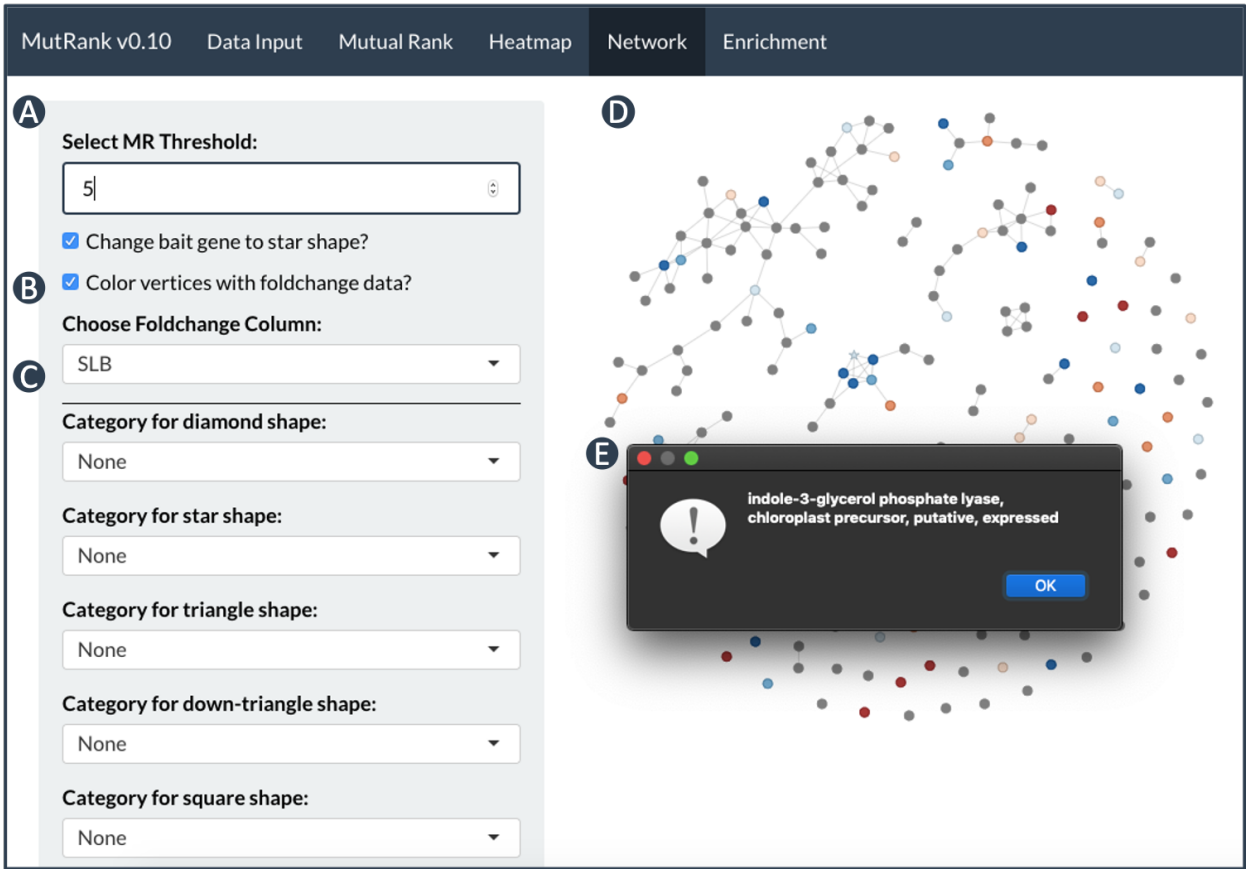
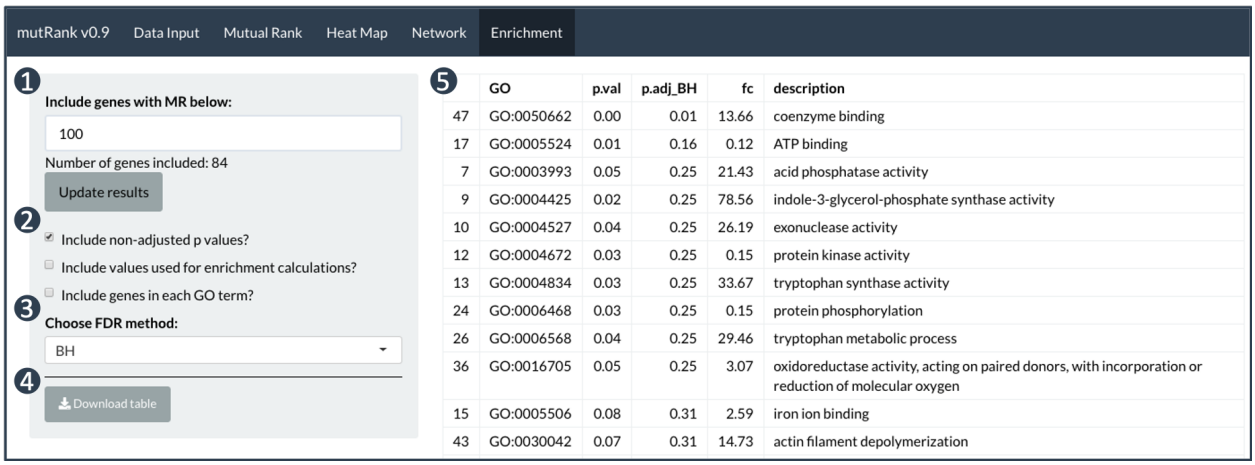


Figure 4: Figure 4: The Network Panel interface of MutRank 1.0 facilitates an initial interrogation of candidate pathway connections. (A) Users define the selected MR threshold for drawing edges between gene vertices in the network and select if the reference gene will star shaped (default) or circle shaped. (B) Users can select a column from the differential expression data to change the colors of the gene nodes according to their expression values. (C) Users can select columns from the custom categories to change the shape of the nodes that match. (D) Output visualization of the dynamic network graph with gene IDs and gene symbols, if available. (E) Users can examine any gene nodes (reference gene selected as an example) to reference available gene annotation(s).

3.5. GO `Enrichment` Tab

Genes that function within similar pathways are more likely to be transcriptionally coregulated. By assigning Gene Ontology (GO) terms to the genes in the MR-based coexpression table we can calculate if certain terms appear more often than expected by random chance. We use the hypergeometric test using the GO databases (MutRank includes basic GO version and the slim plant GO http://geneontology.org/docs/download-ontology/) to calculate the P-values for GO term enrichment. You can choose which method to use to adjust the P-value for false-discvery rate (FDR). To choose which genes will be included in the hypergeometric test we have added an option to select an MR threshold so only genes with MR lower values will be included in the analysis. Below it the total number of included genes for the hypergeometric test will be updated based on updates to the MR-based coexpression table and MR threshold. It is also possible to include in the GO enrichment table the values used for the hypergeometric test for each GO term and the genes assigned to each term. The column names used for the hypergeometric test represent the following: N - Number of genes in the GO annotation files; M - Number of genes annotated with specific GO term; n - Number of included genes from the coexpression table; m - Number of included genes from the coexpression table that are annotated with the specific GO term.



	GO	p.val	p.adj_BH	fc	description
47	GO:0050662	0.00	0.01	13.66	coenzyme binding
17	GO:0005524	0.01	0.16	0.12	ATP binding
7	GO:0003993	0.05	0.25	21.43	acid phosphatase activity
9	GO:0004425	0.02	0.25	78.56	indole-3-glycerol-phosphate synthase activity
10	GO:0004527	0.04	0.25	26.19	exonuclease activity
12	GO:0004672	0.03	0.25	0.15	protein kinase activity
13	GO:0004834	0.03	0.25	33.67	tryptophan synthase activity
24	GO:0006468	0.03	0.25	0.15	protein phosphorylation
26	GO:0006568	0.04	0.25	29.46	tryptophan metabolic process
36	GO:0016705	0.05	0.25	3.07	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
15	GO:0005506	0.08	0.31	2.59	iron ion binding
43	GO:0030042	0.07	0.31	14.73	actin filament depolymerization

Figure 5: GO Enrichment Panel - (A) Use an MR threshold to select which genes from the coexpressiont table will be included in the hypergeometric test for GO term enrichment The total number of genes included will update in the text below the box. (B) Choose which columns to add to the enrichment tables from the values used to calculate the GO term enrichment. (C) Choose which FDR method to use to adjust the P.value. (D) Download the GO term enrichment table to your computer. (E) View the GO term enrichment table results.

If you choose to view the values used to for enrichment calculations,

4. Example Workflows

4.1. Coexpression analysis of the maize benzoxazinoid-biosynthesis pathway

Benzoxazinoids (Bxs) are a highly studied class of maize specialized metabolites involved in defense against herbivores and pathogens.

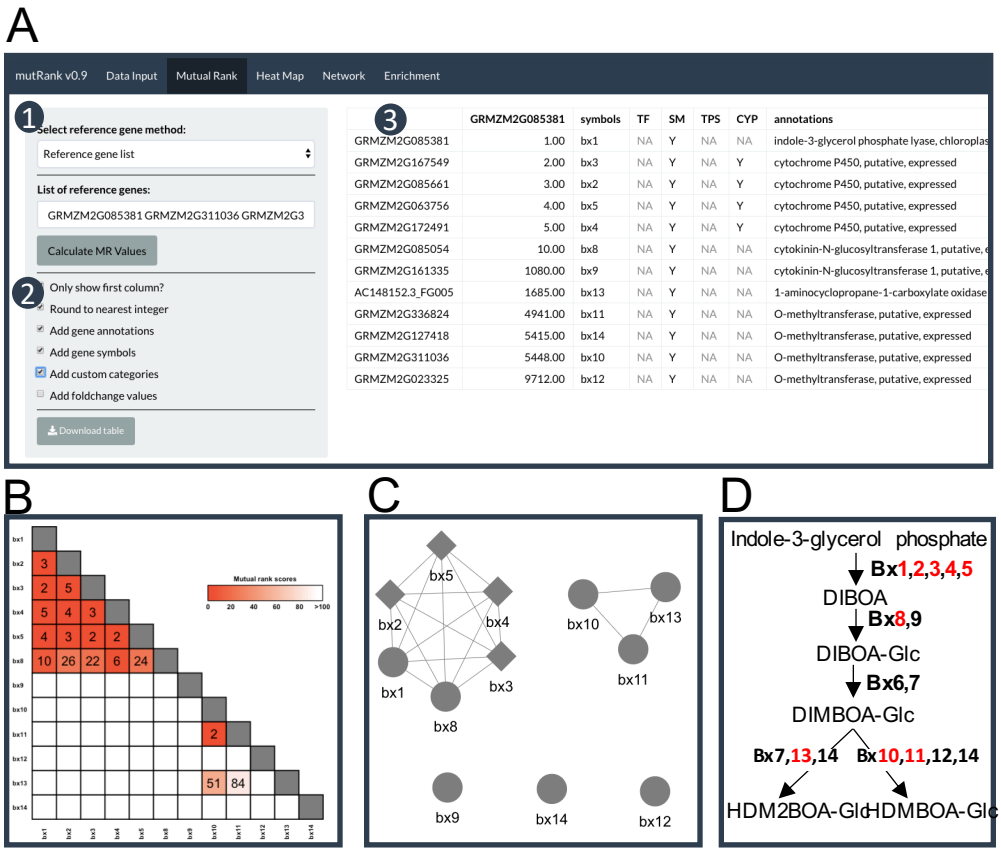


Figure 6: Coexpression analysis of the maize benzoxazinoid-biosynthesis pathway - (A) The Bx biosynthetic pathway includes a series of characterized enzymes (Bx1-14) that were used as a reference gene list (Bx6-7 are not in the expression data and were not included) to calculate the MR values between all the Bxs (1). Users can select the the coexpression data and supporting information (2) that will be integrated and presented in the coexpression table (3). The results of the coexpression analysis can be presented as coexpression heatmap in the Heatmap panel (B) and and coexpression network in the Network panel (C) to show that among the 12 Bxs we included in the reference gene list, using an MR threshold of 100 to draw an edges between genes, we can find 2 clusters of highly-coexpressed Bxs that include 9 of the 14 genes in the original list.

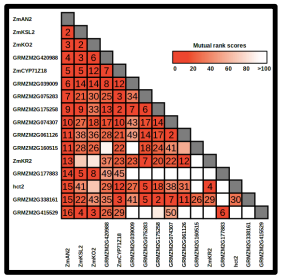
4.2. Coexpression analysis of the maize kauralexin-biosynthesis pathway

Maize specialized metabolites in specific diterpenoid pathways have been implicated in diverse protective roles providing fungal, insect and drought resistance.

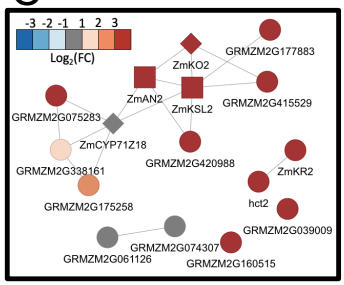
A

	GRMZM2G044481	symbols	TF	SM	TPS	CYP	FC	annotations
GRMZM2G044481	1.00	ZmAN2	NA	Y	Y	NA	11.40	ent-kaurene synthase, chloroplast precursor, putative, expressed
AC214360.3_FG001	2.00	ZmKSL2	NA	Y	Y	NA	7.43	terpene synthase, putative, expressed
GRMZM2G161472	3.00	ZmKO2	NA	Y	NA	Y	6.26	cytochrome P450, putative, expressed
GRMZM2G420988	4.00	NA	NA	NA	NA	NA	6.64	mitochondrial carrier protein, putative, expressed
GRMZM2G067591	5.00	ZmCYP71Z18	NA	Y	NA	Y	NA	cytochrome P450, putative, expressed
GRMZM2G039009	6.00	NA	NA	Y	NA	NA	9.17	SAM dependent carboxyl methyltransferase, putative, expressed
GRMZM2G075283	7.00	NA	NA	Y	NA	NA	10.60	pathogenesis-related Bet v I family protein, putative, expressed
GRMZM2G175258	9.00	NA	NA	NA	NA	NA	2.59	kinase, putative, expressed
GRMZM2G074307	10.00	NA	NA	NA	NA	NA	NA	
GRMZM2G061126	11.00	NA	NA	NA	NA	NA	NA	
GRMZM2G160515	11.00	NA	NA	NA	NA	NA	5.64	
GRMZM2G073929	13.00	ZmKR2	NA	NA	NA	NA	9.36	3-oxo-5-alpha-steroid 4-dehydrogenase, putative, expressed
GRMZM2G177883	14.00	NA	NA	NA	NA	NA	6.65	receptor-like protein kinase 5 precursor, putative, expressed
GRMZM2G114918	15.00	hct2	NA	Y	NA	NA	9.98	transferase family protein, putative, expressed
GRMZM2G338161	15.00	NA	NA	NA	NA	NA	1.66	TKL_IRAK_DUF26-lt.2 - DUF26 kinases have homology to DUF26
GRMZM2G415529	16.00	GRMZM2G415529	NA	Y	NA	NA	10.99	pleiotropic drug resistance protein, putative, expressed

B



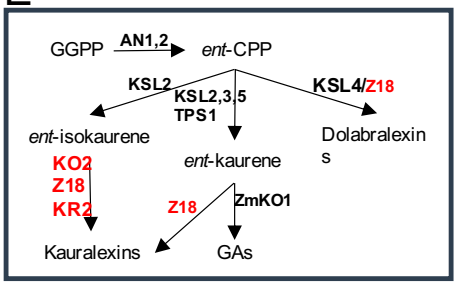
C



D

GO	p.val	p.adj_BH	fc	description
GO:0006952	0.00	0.01	18.25	defense response
GO:0042742	0.00	0.02	63.86	defense response to bacterium
GO:0009607	0.00	0.03	35.48	response to biotic stimulus
GO:0010333	0.00	0.03	12.60	terpene synthase activity
GO:0050832	0.00	0.03	45.62	defense response to fungus
GO:0005506	0.00	0.03	3.51	iron ion binding
GO:0005515	0.00	0.04	0.28	protein binding
GO:0016705	0.00	0.04	3.64	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0016829	0.01	0.04	8.71	lyase activity

E



]

5. Acknowledgements

We would also like to thank the contributors of the cited R packages.

6. License