MutRank

Elly Poretsky, Alisa Huffaker

April 6, 2020

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 within similar pathways are more likely to be transcriptionally coregulated and can be used to predict functional associations and putative gene function.

Many databases and webtools have been developed to facilitate coexpression analyses and many of them use the 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 specialized maize metabolic enzymes within 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 Data Preparation

The MutRank folder contains a separate folder for each of the supported data type. Files in these folders will be automatically included in the dropdown menu for data input field and loaded once selected. We also included an option to upload files manually (this was included for running MutRunk on as Shiny Server instance where users might not have access to the folders). Once you have selected the files to load you can press the `Save Default` button so the same files will be automatically loaded next time you start MutRank (default settings are saved in the main folder in `default_files.csv`).

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

3. Navigating the MutRank Tabs

3.1 Main `Data Input` Tab

Data Input is the first tab to appear when you start MutRank. MutRank only requires expression data to preform basic coexpression analyses and will integrate any supporting information to facilitate prediction of putative gene functions. If the expression data and supporting information files are correctly formatted that will appear in the appropriate selection box in the Data Input tab (Fig. 1), make sure that the files have the correct suffix and file format. Looking at the example files might also be helpful at solving certain problems.

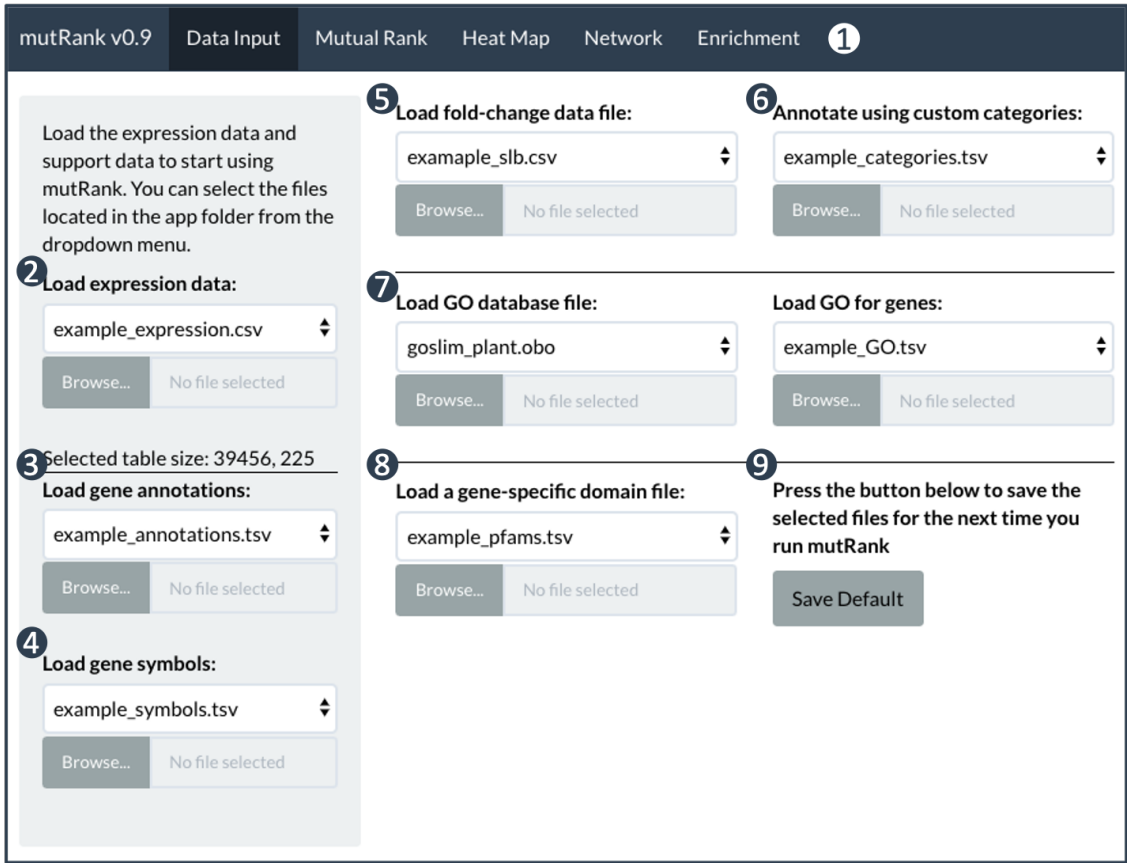
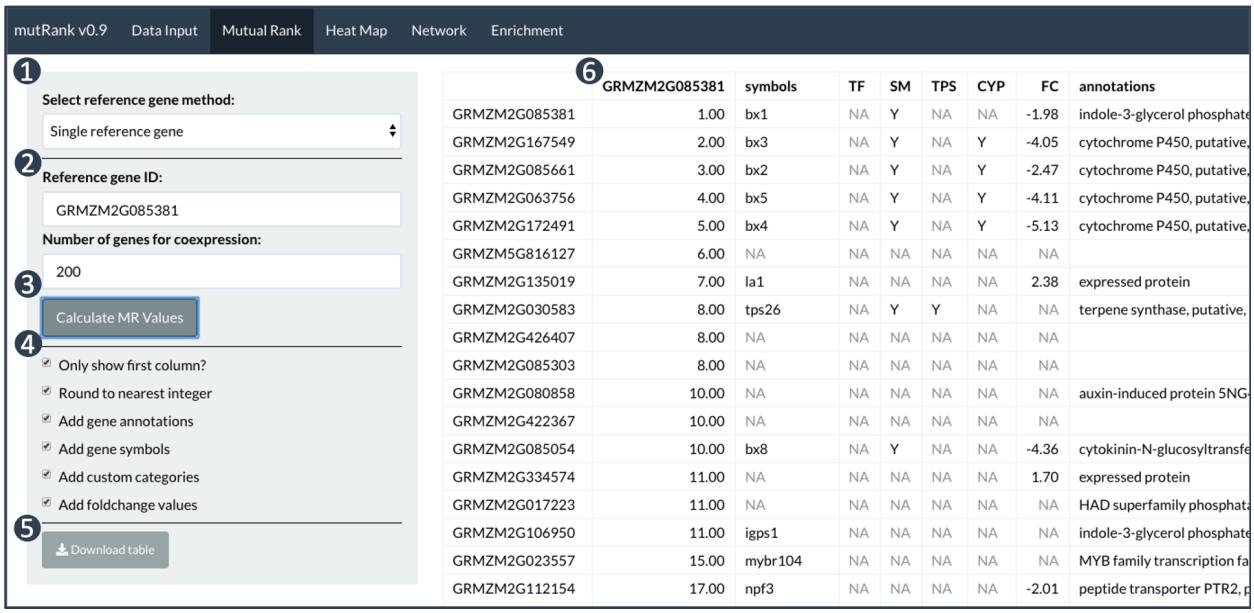


Figure 1: Screenshot of the main tab for Data Input. You can navigate between the different components of MutRank using the top tab panel (1). In the left side panel you can load expression data (2), gene annotations (3) and gene symbols (4). In the main panel you can load the differential expression data (5), custom categories (6), GO database and GO assignments (7), protein domain assignments (8) and save default files to open on next MutRank run (9).

3.2 `Mutual Rank` Tab

The Mutual Rank tab contains the main interface for calculating the MR-based coexpression table using user-provided expression data and supporting information files.



	GRMZM2G085381	symbols	TF	SM	TPS	CYP	FC	annotations
GRMZM2G085381	1.00	bx1	NA	Y	NA	NA	-1.98	indole-3-glycerol phosphate
GRMZM2G167549	2.00	bx3	NA	Y	NA	Y	-4.05	cytochrome P450, putative,
GRMZM2G085661	3.00	bx2	NA	Y	NA	Y	-2.47	cytochrome P450, putative,
GRMZM2G063756	4.00	bx5	NA	Y	NA	Y	-4.11	cytochrome P450, putative,
GRMZM2G172491	5.00	bx4	NA	Y	NA	Y	-5.13	cytochrome P450, putative,
GRMZM5G816127	6.00	NA	NA	NA	NA	NA	NA	
GRMZM2G135019	7.00	la1	NA	NA	NA	NA	2.38	expressed protein
GRMZM2G030583	8.00	tps26	NA	Y	Y	NA	NA	terpene synthase, putative,
GRMZM2G426407	8.00	NA	NA	NA	NA	NA	NA	
GRMZM2G085303	8.00	NA	NA	NA	NA	NA	NA	
GRMZM2G080858	10.00	NA	NA	NA	NA	NA	NA	auxin-induced protein 5NG
GRMZM2G422367	10.00	NA	NA	NA	NA	NA	NA	
GRMZM2G085054	10.00	bx8	NA	Y	NA	NA	-4.36	cytokinin-N-glucosyltransfe
GRMZM2G334574	11.00	NA	NA	NA	NA	NA	1.70	expressed protein
GRMZM2G017223	11.00	NA	NA	NA	NA	NA	NA	HAD superfamily phosphat
GRMZM2G106950	11.00	igps1	NA	NA	NA	NA	NA	indole-3-glycerol phosphate
GRMZM2G023557	15.00	mybr104	NA	NA	NA	NA	NA	MYB family transcription fa
GRMZM2G112154	17.00	npf3	NA	NA	NA	NA	-2.01	peptide transporter PTR2, p

Figure 2: Screenshot of the Mutual Rank tab. Start by selecting one of the three reference-gene methods (1). Insert gene(s) in the appropriate box and how many from the top coexpressed genes (based on PCC) to include in the analysis (2). To start calculating the MR coexpression table press the Calculate MR Values button (3). Select which of the options and supporting information to integrate with the coexpression table (4). You can download the coexpression table (5) as you see it (6).

3.3 Coexpression `Heatmap` Tab

The coexpression heatmap is generated using ggplot2. Use the side panel to

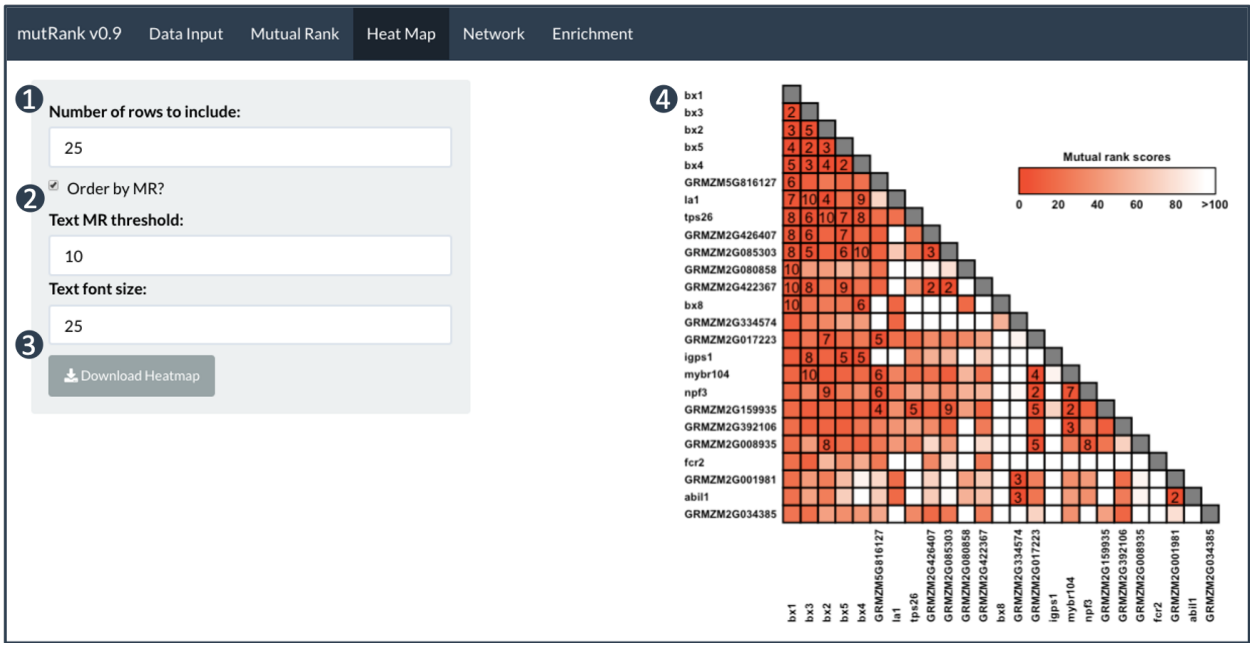


Figure 3: Heatmap Panel - (A)

3.4 Coexpression `Network` Tab

The coexpression network is created using igraph and converted to a dynamic java-script netowrk using vizNetwork package. Use the side panel to edit the final network to show specified aspects of the supporting information.

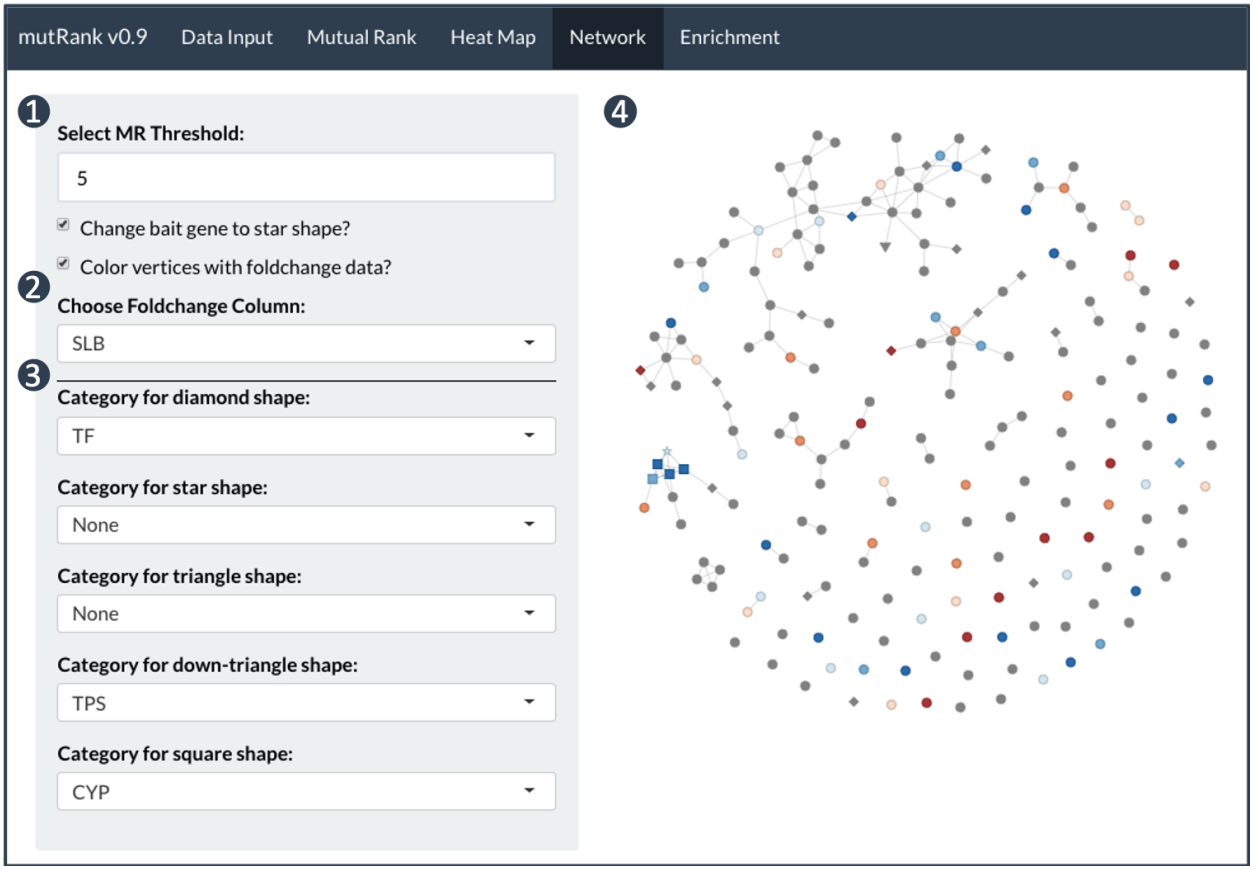
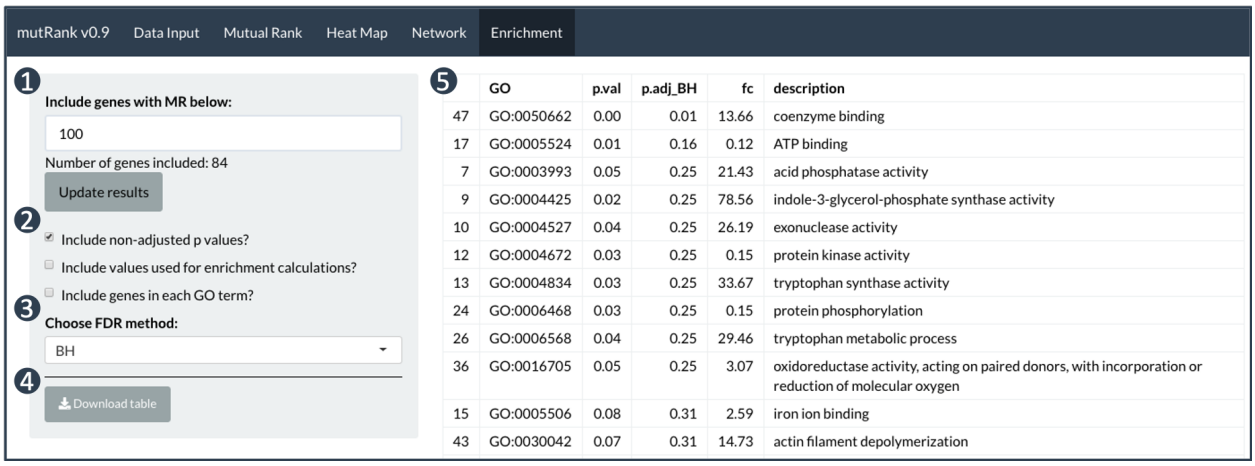


Figure 4: Network Panel - (A) Select MR threshold for drawing edges between vertices.

3.5 GO `Enrichment` Tab

GO enrichment is calculated using a hypergeometric test.



	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) Select MR threshold to include genes in the GO enrichment calculations.

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.

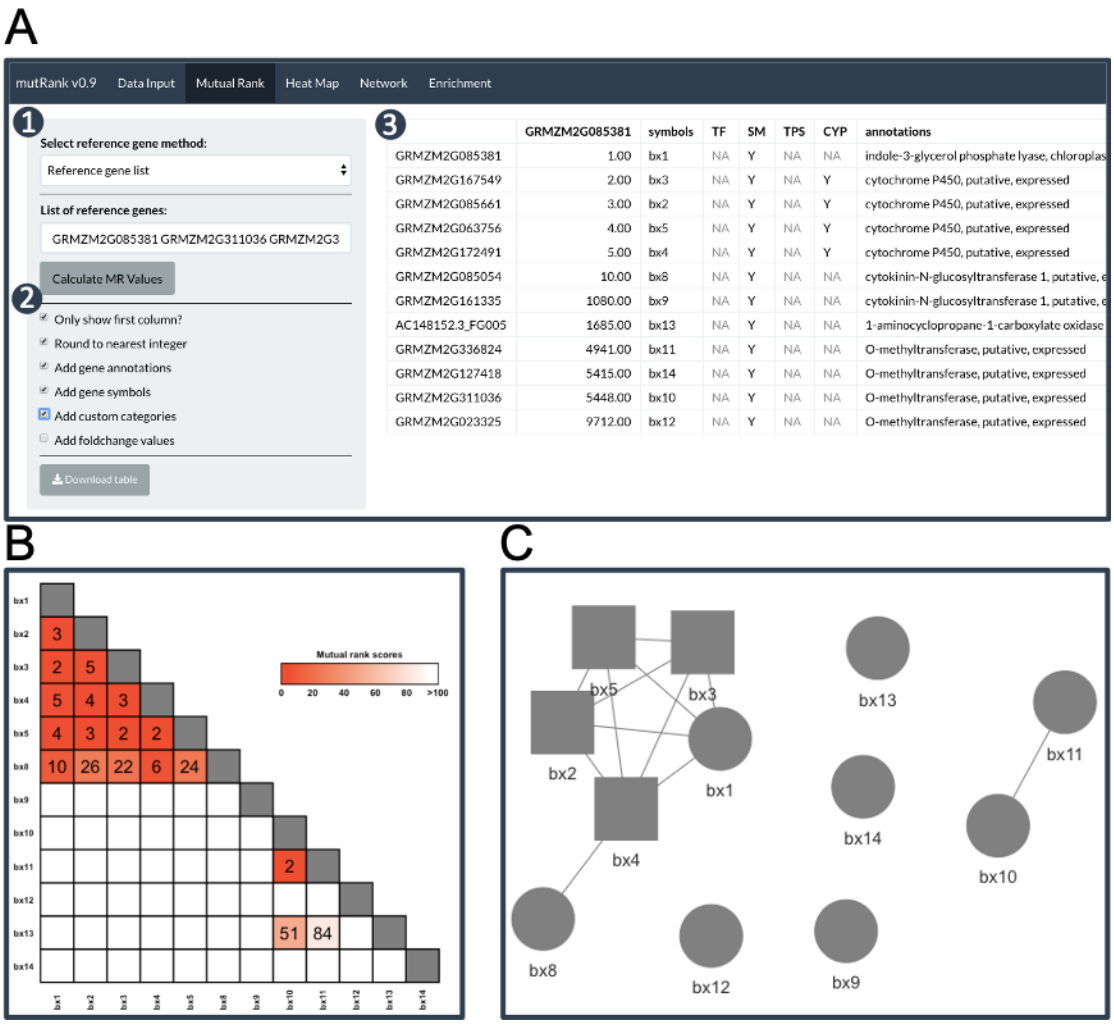


Figure 6: Coexpression analysis of the maize benzoxazinoid-biosynthesis pathway - (A) The Bx biosynthetic pathway includes a series of characterized enzymes (Bx1-14) that were used as a reference gene list (Bx6-7 are not in the expression data and were not included) to calculate the MR values between 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.

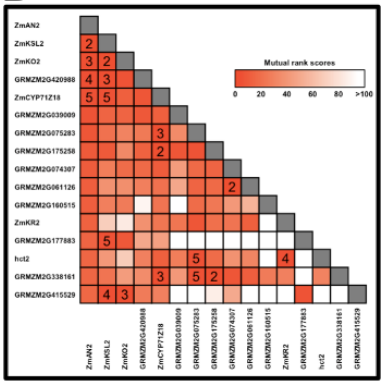
4.2 Coexpression analysis of the maize kauralexin-biosynthesis pathway

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

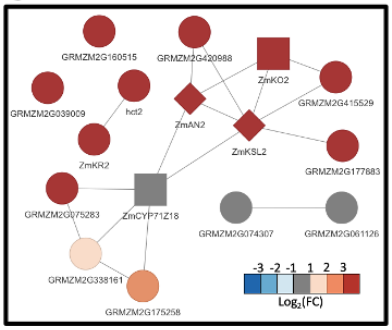
A

	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-lf.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

Figure 7: Coexpression analysis of the maize kauralexin-biosynthesis pathway - (A) Characterizing of the the maize kauralexin-biosynthesis pathway leveraged targeted MR-based coexpression analyses. We used ZmAN2, one of the known genes responsible for catalyzing a key pathway precursor, as a single reference gene to calculate the MR values for the top 200 most highly coexpressed genes (based on PCC) and integrated the results with the supporting information. We selected the top 15 most highly coexpressed genes (based on MR) in the table to generate (B) a coexpression heatmap and (C) a coexpression network figure to show a cluster of coexpressed TPSs (ZmAN2 and ZmKSL2, diamond shape vertices) and CYPs (ZmKO2 and ZmCYP71Z18, square shaped vertices), their induced expression 24 hours after SLB treatment. (D) GO enrichment analysis of genes with MR<100 included 124 genes. Term enrichment P-values were calculated with a hypergeometric test using the full GO database and P-values were adjusted using the Bonferroni–Holm method. GO enrichment results indicate that the genes coexpressed with ZmAN2 are involved in biotic stress responses.

5. Acknowledgements

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

6. License

MutRank is available under the...