NASH-Predictor V1.0 Documentation

1. Installation

1.1 Installation method:

The current version of NASH-Predictor is a green software. For software compiled in R, users only need to download the https://github.com/sirpan/NASHnet/tree/main/software file in github and use Rstudio to run the run.R file line by line

1.2 Operating environment

1.2.1 Hardware environment
- ✓ Minimum configuration: If you only need to study transcriptome data with a small number of genes (selectable number) and a small number of samples, you only need a 64-bit dual-core processor and 4 GB of memory.
- ✓ Recommended: If you need to use a large sample and a large number of genes, 32 GB or 64 GB of RAM is recommended
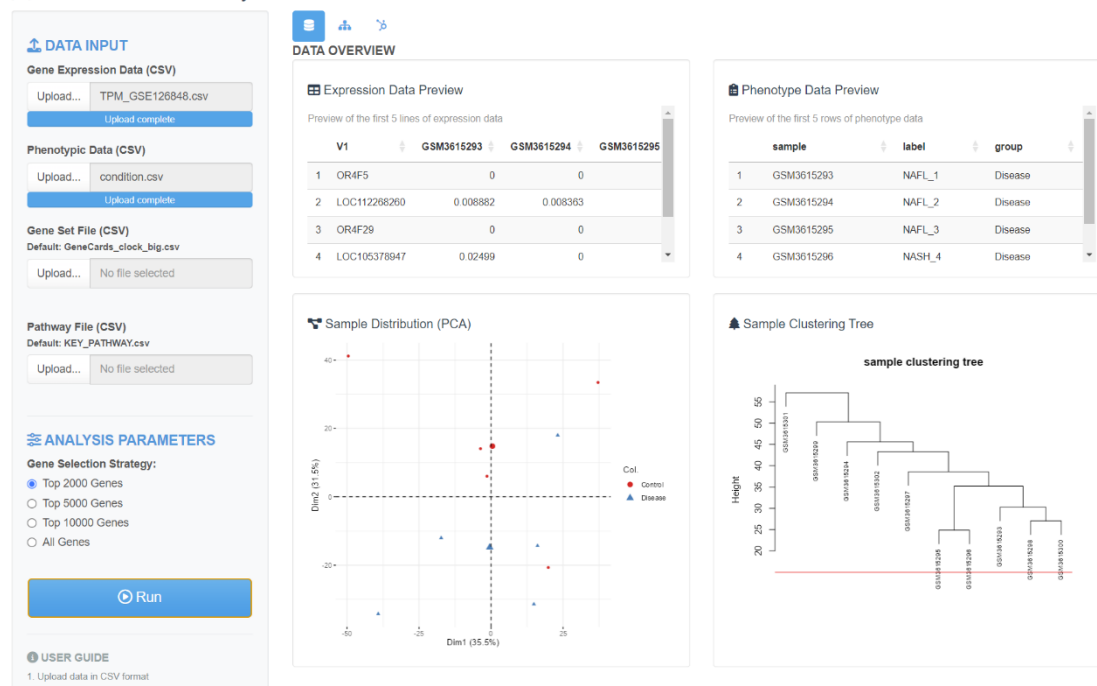
1.2.2 Software environment
- ✓ Operating system: CentOS 7 or a compatible Linux operating system or Windows 10 or later
- ✓ Software dependencies: Anaconda or Miniconda and Rstudio > version 3.0.

2. Interface introduction

2.1 Input interface

On the left, you must input **transcriptome data or proteomic** data in the first column, and enter its status, **disease/control** in the second column (**refer to the example in the test folder in the current directory TPM_ GSE126848.csv files and condition.csv files**), while the data in the third and fourth columns are not mandatory to enter data, you can customize the data, y**ou can enter the gene set and pathway of interest by yourself, or you can use the default data (refer to the example: the test folder in the current directory key_target_set.csv the file and KEY_ PATHWAY.csv the file)**, you can choose to filter the gene set according to your computer configuration, the default is all, and then click Run to run the NASH-Predictor software with one click. On the left, we have an interface that previews the dataset input, and after the run is complete, two plots are generated below about the sample analysis:

✓ **The Sample Distribution (PCA) diagram** is a PCA dimensionality reduction analysis based on the input data of the sample, and the origin and triangle in the graph represent the spatial distribution of the control group and the disease group, respectively

✓ **The Sample Clustering Tree diagram** is a calculated correlation treemap of samples, the farther away the samples in the diagram are, the less similar the samples are, if the samples are all within the red line, it means that the samples do not need to be deleted, and if the red line is exceeded, it means that the samples are deleted during the calculation

Sample visual clustering analysis can provide users with a visual representation of the quality of their input data, and users can judge whether the results are reliable through two graphs (if the clustering difference between samples is large, the module analysis results are inaccurate)

2.2 Key module results interface
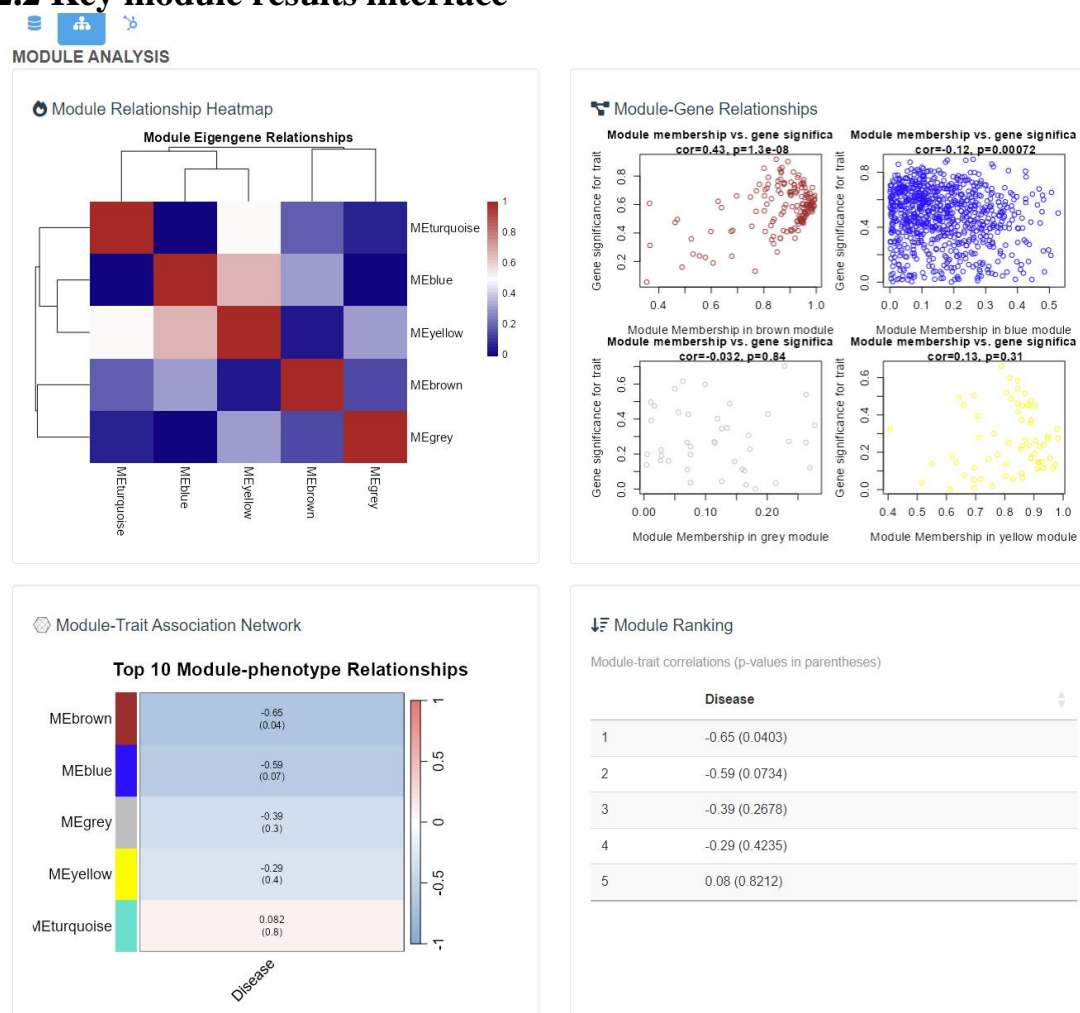
The analysis of key modules is divided into four parts:

✓ **Module Correlation Analysis**: The first heatmap represents the module correlation analysis, and the darker the color, the more correlation

✓ **Correlation analysis of modules and intra-module genes**: Represented by a scatter plot, the correlation between the top 4 modules with the highest correlation and internal genes is shown here.

✓ **The darker the color represents the more related to the disease**, the number in the bracket represents the p-value value, and the color on the left is the name of the module.

✓ **Module Ranking Table:** The module names on the inside and left correspond to their relevance to the disease

This section facilitates the visualization of the results of WGCNA's discovery of key modules for comparison by the user

2.3 Scoring of key genes

PPI NETWORK

NASH-Score Results Panel

Search:

Key targets and composite scores

Gene ID	Gene Symbol	Finder Score	Pathway Match	Total Score
7514	XPO1	0.884	0.250	1.000
5962	RDX	0.314	0.500	0.579
2317	FLNB	0.271	0.500	0.536
2	A2M	0.796	1.000	0.508
1982	EIF4G2	0.219	1.000	0.478
1956	EGFR	0.799	0.208	0.370
821	CANX	0.774	0.400	0.266
80854	SETD7	0.275	0.333	0.262
1500	CTNND1	0.258	0.333	0.246
10769	PLK2	0.671	1.000	0.242

Showing 1 to 10 of 819 entries   Previous 1 2 3 4 5 … 82 Next

After the key modules are discovered in WGCNA, the software will automatically extract the genes in the top two modules in terms of relevance, and then construct a protein-protein interaction network and use the built-in algorithm to score the targets

The table provides a search function, which mainly shows the Finder Score (calculated by the Finder method), the Pathway Match score and the pathway match you provide, and the total score

4

3. Download the results

All results will be automatically downloaded at the end of the run, for example, the scoring **results of the target score table above will be available in the NASHpredictor\inst\app\NASH_Score folder**