**Supplementary material – Example Pipelines**

The objective of this supplementary material is to present the workflow behind each type of the results user want to get. Users can discover many pipelines in TSUNAMI. In this supplementary material, 4 examples are given. The most comprehensive one is in the first section, then the direct gene ontology (GO) enrichment analysis pipeline and survival analysis pipeline followed. Finally, another option allows user to draw functional plots (circos plot) directly.

1. **Comprehensive gene co-expression and gene ontology enrichment analysis pipeline**

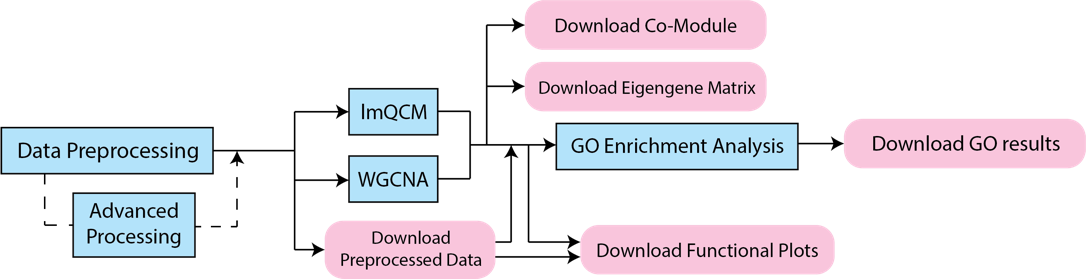


Fig. S1

The most canonical case is the comprehensive pipeline (Fig. S1), which including data preprocessing (optional: advanced processing for survival analysis, see Survival Analysis Pipeline for details), weighted network co-expression analysis (two algorithms lmQCM and WGCNA available), GO enrichment analysis, and optional downloadable files.

In this pipeline, user can select series matrices as raw data from NCBI GEO database, or upload raw data with csv, txt, xls, or xlsx formats. By the time user finish the data preprocessing/advanced processing, user can continue on co-expression analysis, or download preprocessed data in csv or txt format. The preprocessed data can also use for circos plot.

After co-expression analysis, thousands of genes will be clustered into several clusters if they appear to have co-expression pattern, sorted by length of clusters (number of genes in each of them). Note that the results depend on the parameter chosen in lmQCM or WGCNA algorithms. User can then download co-module and eigengene matrix in csv or txt formats.

Finally, GO enrichment analysis based on Enrichr [1] [2] provides 14 GO enrichment results to users: (1) Biological Process; (2) Molecular Function; (3) Cellular Component; (4) Jensen DISEASES; (5) Reactome; (6) KEGG; (7) Transcription Factor PPIs; (8) Genome Browser PWMs; (9) TRANSFAC and JASPAR PWMs; (10) ENCODE TF ChIP-seq; (11) Chromosome Location (Cytoband); (12) miRTarBase; (13) TargetScan microRNA; (14) ChEA.

1. **Direct gene ontology enrichment analysis pipeline**

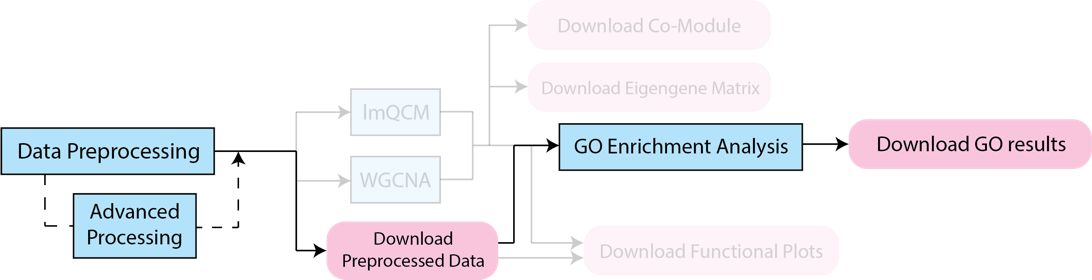


Fig. S2

Another example pipeline is simply GO analysis right after cleaning the raw data, as shown in Fig. S2 (methods/analysis with 20% transparency will not be performed). This pipeline allows user to perform GO analysis on entire gene sets without any co-expression analysis.

Through this pipeline, users would found TSUNAMI provides a powerful data cleaning/preprocessing solution, bundled with a GO enrichment analysis. Data preprocessing and advanced processing are also optional by not performing any data preprocessing tasks, if users want to preserve raw data all the way to GO enrichment analysis.

1. **Survival Analysis pipeline**

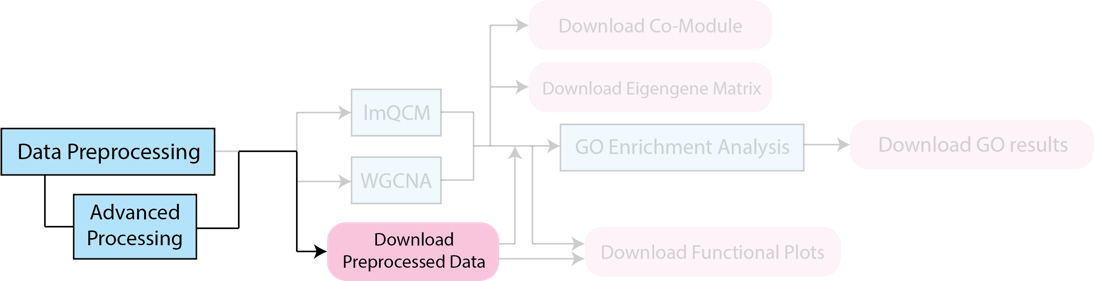


Fig. S3

Survival analysis is also an interesting pipeline provided by TSUNAMI. Users have to upload their own file containing numerical value of OS (overall survival) Indicator/group and OS value, or numerical value of EFS (event-free survival) Indicator/group and EFS value. Failure to upload the data with the requirement or OS/EFS values are non-numerical may not produce results. Fig. S3.1 shows an example with screenshots of data uploading (a) and advanced processing (b) procedures.

“survdiff” function from R package “survival” is adopted to calculate the Chi-square test p-value. For example, survival analysis performs on 11 samples OS group and value. Then based on sorted (increasingly) samples along with their OS group, lower 5 samples will test against higher 6 samples to get the p-value.

Finally, the results are given in “3. Choose Method” Section “Verify Final Data” Panel (Fig. S3.2 (a)). The corresponding p-values are shown in the second column. Moreover, it is of users interest to generate circos plot (Fig. S3.2 (b), plot contains 123 genes) based on all final genes by clicking the yellow button “Circos Plot for All Genes” in Fig. S3.2 (a).

|  |  |
| --- | --- |
| ../Screen%20Shot%202018-04-01%20at%205.19.55%20PM.png | ../Screen%20Shot%202018-04-01%20at%207.01.16%20PM.png |
| (a) | (b) |

Fig. S3.1

|  |  |
| --- | --- |
| ../Screen%20Shot%202018-04-01%20at%207.03.41%20PM.png | ../Screen%20Shot%202018-04-01%20at%207.05.46%20PM.png |
| (a) | (b) |

Fig. S3.2

1. **Generate circos plot based on a group of genes**

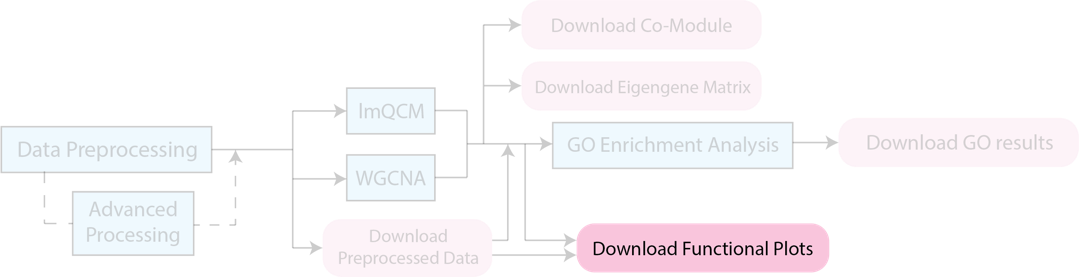


Fig. S4

Circos plot is a very useful plot for visualizing the positions of genes on chromosomes. In TSUNAMI, we support human genomes hg38 (GRCh38) and hg19 (GRCh37). To match the gene symbol to chromosomes’ starting and ending site, we use reference Gene table from UCSC genome browser [3]. If multiple starting/ending site are matched, we choose the longest one with calculated by:

TSUNAMI uses R package “circlize” [4] to draw the circos plot. To get the circos plot, user can either directly input gene symbols separated by carriage return character (A.K.A. “\n”) in the format as shown in Fig. S4.1, or use the results provided by other pipelines. By updating the plots, users can also choose the size of the plots and decide whether gene symbols and links between genes should be shown on the graph.

An example output of circos plot in Fig. S4.2 is of human genome GRCh38 (hg38) with links between genes, indicated by gene symbols.



Fig. S4.1

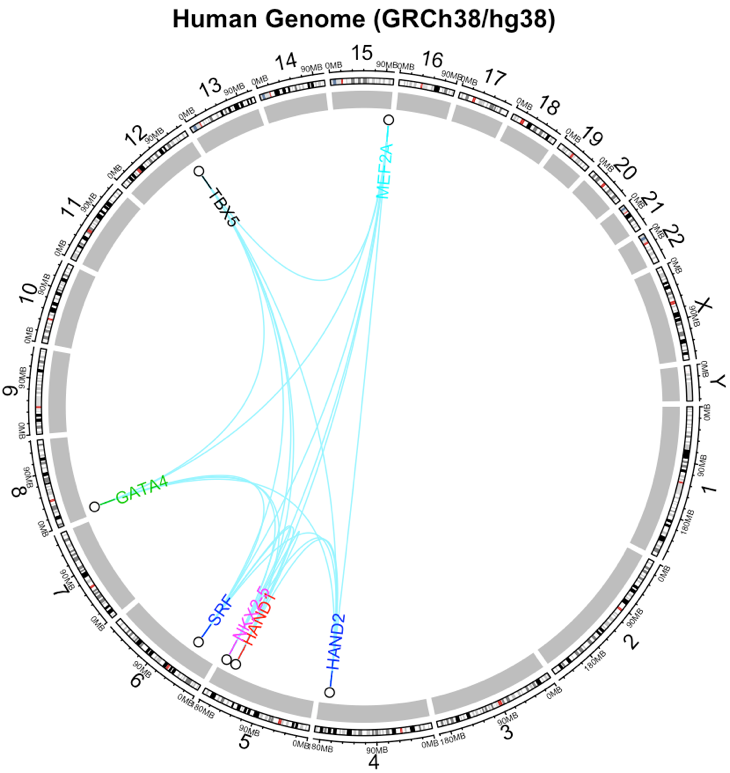


Fig. S4.2

**Reference**

[1] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., ... & Ma’ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics, 14(1), 128.

[2] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., ... & McDermott, M. G. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research, 44(W1), W90-W97.

[3] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. Genome research, 12(6), 996-1006.

[4] Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize implements and enhances circular visualization in R. Bioinformatics, 30(19), 2811-2812.