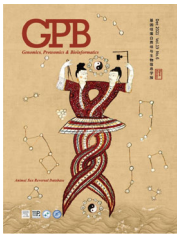
Genomics Proteomics Bioinformatics 19 (2021) 1023–1031

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



APPLICATION NOTE

TSUNAMI: Translational Bioinformatics Tool Suite for Network Analysis and Mining

Zhi Huang[1,2,#], Zhi Han[3,#], Tongxin Wang[4], Wei Shao[3], Shunian Xiang[5], Paul Salama[2], Maher Rizkalla[2], Kun Huang[3,*,§], Jie Zhang[5,*]

[1]*School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA*
[2]*Department of Electrical and Computer Engineering, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA*
[3]*Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[4]*Department of Computer Science, Indiana University, Bloomington, IN 47405, USA*
[5]*Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*

Received 16 October 2018; revised 3 April 2019; accepted 31 May 2019
Available online 8 March 2021

Handled by Edwin Wang

Abstract Gene co-expression network (GCN) mining identifies gene modules with highly correlated expression profiles across samples/conditions. It enables researchers to discover latent gene/molecule interactions, identify novel gene functions, and extract molecular features from certain disease/condition groups, thus helping to identify disease biomarkers. However, there lacks an easy-to-use tool package for users to mine GCN modules that are relatively small in size with tightly connected genes that can be convenient for downstream gene set enrichment analysis, as well as modules that may share common members. To address this need, we developed an online GCN mining tool package: TSUNAMI (Tools SUite for Network Analysis and MIning). TSUNAMI incorporates our state-of-the-art lmQCM algorithm to mine GCN modules for both public and user-input data (microarray, RNA-seq, or any other numerical omics data), and then performs downstream gene set enrichment analysis for the identified modules. It has several features and advantages: 1) a user-friendly interface and real-time co-expression network mining through a web server; 2) direct access and search of NCBI Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases, as well as user-input gene expression matrices for GCN module mining; 3) multiple co-expression analysis tools to choose from, all of which are highly flexible in regards to parameter selection options; 4) identified GCN modules are summarized to eigengenes, which are convenient for users to check their correlation with other clinical traits; 5) integrated downstream Enrichr enrichment analysis and links to other gene set enrichment tools; and 6) visualization of gene loci by Circos plot in any step of the process. The web service is freely accessible through URL: https://biolearns.medicine.iu.edu/. Source code is available at https://github.com/huangzhii/TSUNAMI/.

KEYWORDS Network mining; Gene co-expression network; Transcriptomic data analysis; lmQCM; Web server; Survival analysis

Introduction

Gene co-expression network (GCN) mining is a popular bioinformatics approach to identify densely connected gene

*Corresponding authors.
E-mail: jizhan@iu.edu (Zhang J), kunhuang@iu.edu (Huang K).
[#]Equal contribution.
[§]Current address: Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
https://doi.org/10.1016/j.gpb.2019.05.006

modules, which are linked by their highly correlated expression profiles. It helps biologists discover latent gene/molecule interactions and identify novel gene functions, disease pathways, biomarkers, and insights for disease mechanisms. GCN mining approaches such as WGCNA [1] and lmQCM [2] have been increasingly used [3–7]. Compared to the more popularly used WGCNA package, lmQCM is capable of mining smaller densely connected GCN modules. It also allows overlapping membership in the output modules. Such features are more consistent with biological networks in which the same genes may participate in multiple pathways, where a small group of genes are more likely to be synergistically regulated in local pathway functions. In addition, gene modules with smaller size derived from lmQCM usually generate more meaningful gene set enrichment results, which have been successfully applied to many diseases and cancer types [8–17].

Currently, several online databases exist that curate transcriptomic data. For instance, PanglaoDB (https://panglaodb.se/) collects single-cell RNA sequencing (scRNA-seq) data from mice and humans; scRNASeqDB [18] provides an scRNA-seq database for gene expression profiling in humans; *recount2* [19] provides publicly available analysis-ready gene and exon counts datasets. However, all of these databases focus on data collection and curation. To the best of our knowledge, there is no tool offering the complete pipeline that can directly process transcriptomic data, mine GCN modules, carry out gene set enrichment analysis, and provide visualization for the results. To meet such needs, we implemented our web-based analysis tool suite Tools SUite for Network Analysis and MIning (TSUNAMI).

For users' convenience, mRNA-seq data from The Cancer Genome Atlas (TCGA; Illumina HiSeq RSEM genes normalized from https://gdac.broadinstitute.org/) and NCBI Gene Expression Omnibus (GEO) are directly incorporated into TSUNAMI. GEO hosts a large number of transcriptomic datasets generated from multiple platforms, including microarray and RNA-seq data. Other data types, such as miRNA-seq and DNA methylation, are also compatible with TSUNAMI. In fact, TSUNAMI can handle any numerical matrix data regardless of the omics data type. TSUNAMI not only incorporates the newly released lmQCM algorithm, but also includes the WGCNA package for users to explore and compare GCN modules generated from two different algorithms. We offer highly flexible parameter choices in each step to users who want to fine tune each algorithm to suit their own data and goal.

Prior to data mining, a data pre-processing interface has been designed to address differences in the input data formats and to filter the data in order to remove noise for GCN mining. Each step of pre-processing is transparent to users and can be adjusted according to their preferences and needs.

Furthermore, our website directly incorporates enrichment analysis of the gene modules and Circos plot function for researchers to explore the enriched biological terms and gene locations in the output GCN modules. It also provides a tool for survival analysis with respect to each GCN module's eigengene values. All the aforementioned functions only require button clicks from users. The design of such a user-friendly interface in our TSUNAMI pipeline provides a one-stop comprehensive analysis tool suite for biological researchers and clinicians to perform transcriptomic data analyses without any programming skill or data mining knowledge.

Method

A flowchart of the TSUNAMI pipeline is presented in **Figure 1**. The entire pipeline is implemented in R language with Shiny server pages. In the future, it will be upgraded

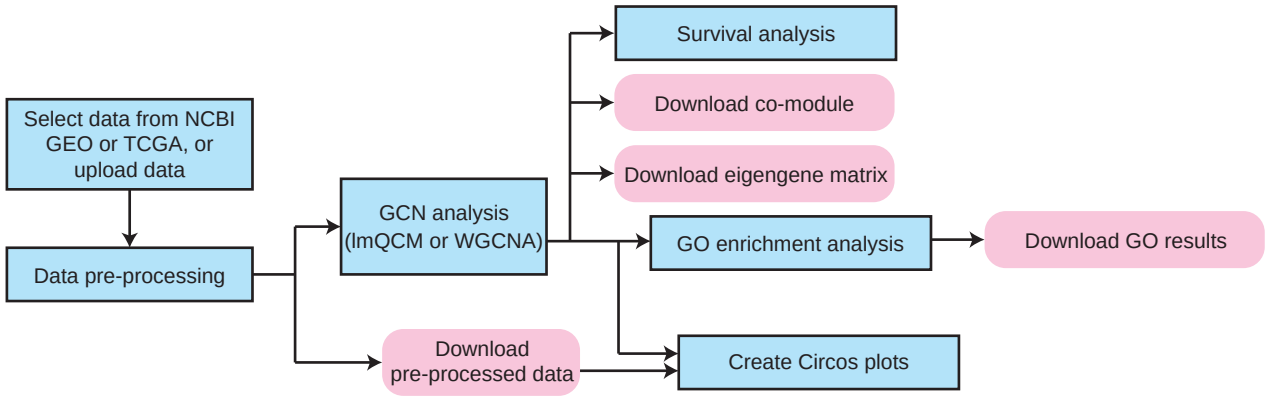


Figure 1 Flowchart of TSUNAMI
In this flowchart for TSUNAMI pipeline, blue rectangles represent pipeline operations; rounded rectangles in pink represent download processes. TSUNAMI, Tools SUite for Network Analysis and Mining; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas; GCN, gene co-expression network; GO, Gene Ontology.

with Python to improve computing efficiency in the module mining step. Some front-end interfaces and functions are implemented using JavaScript. With TSUNAMI, users can choose to use multiple types of data formats, including TCGA RNA-seq data, gene expression microarray data from GEO (in the format of GSE series matrix data), RNA-seq data from GEO, and user-defined numerical matrix data (such as microarray, RNA-seq, scRNA-seq, and DNA methylation data). Instead of searching the GEO database manually, TSUNAMI provides a friendly interface for users to retrieve data from GEO by utilizing keywords and offers a flexible selection tool to retrieve a relevant GSE dataset to perform GCN analysis. Users can also choose a specific omics data type on the GEO database if keywords are entered in the search window to indicate the desired data type. In our testing, only a smaller portion of GSE data was not able to be processed (*e.g.*, 12 out of first 1000 GSE data), most of which were legacy microarray data that contain too much missing data or too small of a sample size. On the website, a variety of example datasets ranging from microarray to scRNA-seq data are listed on TSUNAMI for users' reference. TSUNAMI also provides an upload bar for users to upload local files in various formats (*e.g.*, CSV, TSV, XLSX, and TXT). The data uploading interface is shown in **Figure 2**A. In this study, one microarray dataset (GSE17537 from GEO) was chosen as an example to demonstrate the features of TSUNAMI. GSE17537 contains gene expression data of 55 colorectal cancer patients from the Vanderbilt Medical Center (VMC) generated from the Affymetrix HU133 2.0 Plus Genechip with 54,675 probesets [20,21].

Results

Online data pre-processing

One issue of the microarray dataset from GEO is that different platforms adopt different rules when converting probeset IDs to gene symbols. To make this step easier for users, probeset IDs in GSE data matrix from GEO can be converted to gene symbols using R package "BiocGenerics" [22] by only one click. For instance, for the GSE17537 dataset, the annotation platform is GPL570. TSUNAMI then automatically identifies the annotation platforms of the data from GEO. During the conversion, TSUNAMI 1) removes rows with empty gene symbols and 2) selects the rows with the largest mean expression value when multiple probesets are matched to the same gene symbol. The user interface of the data pre-processing step is shown in Figure 2B.

Additional data filtering steps include: 1) converting "NA" value (not a number value) to 0 in expression data, to

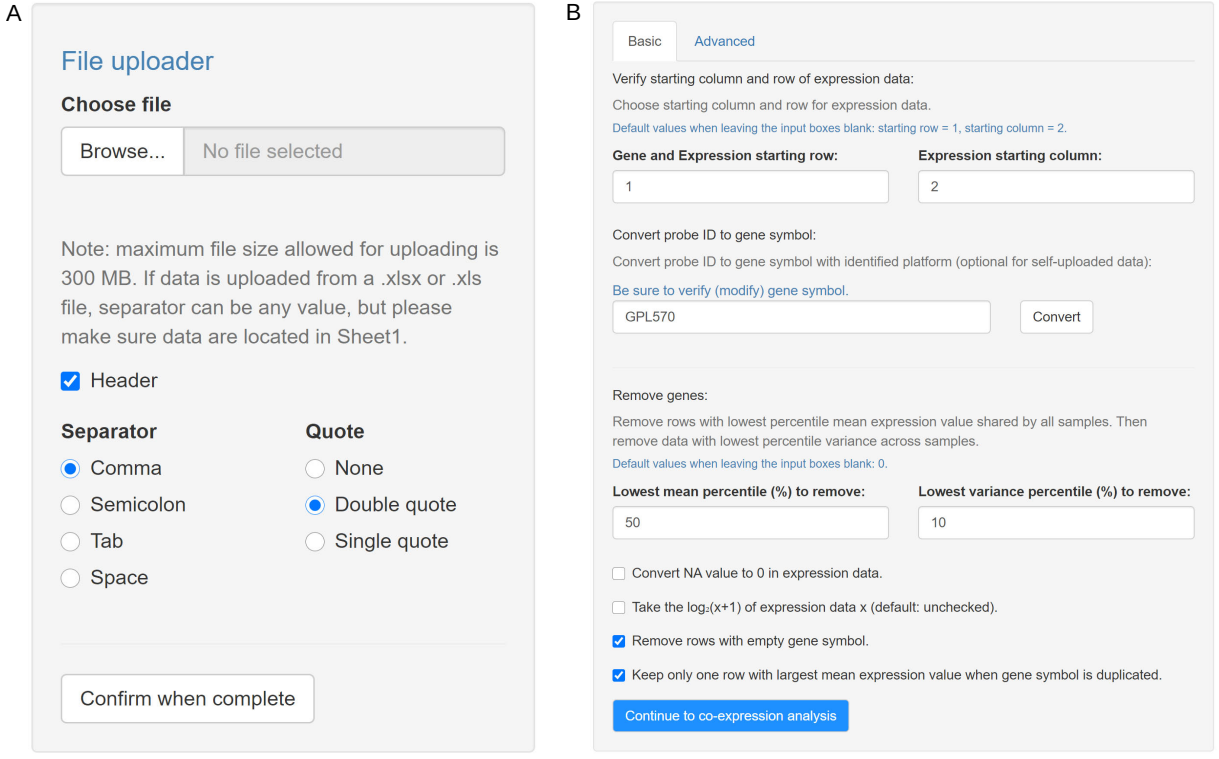


Figure 2 **Dataset selection and the data pre-processing panel**
A. Data can be uploaded manually or chosen from the NCBI GEO database (not shown in the figure). When uploading the data, the maximum file size that TSUNAMI allows is 300 megabytes. Header, separators, and quote methods can be adjusted by users. **B.** The data pre-processing panel includes several pre-processing steps.

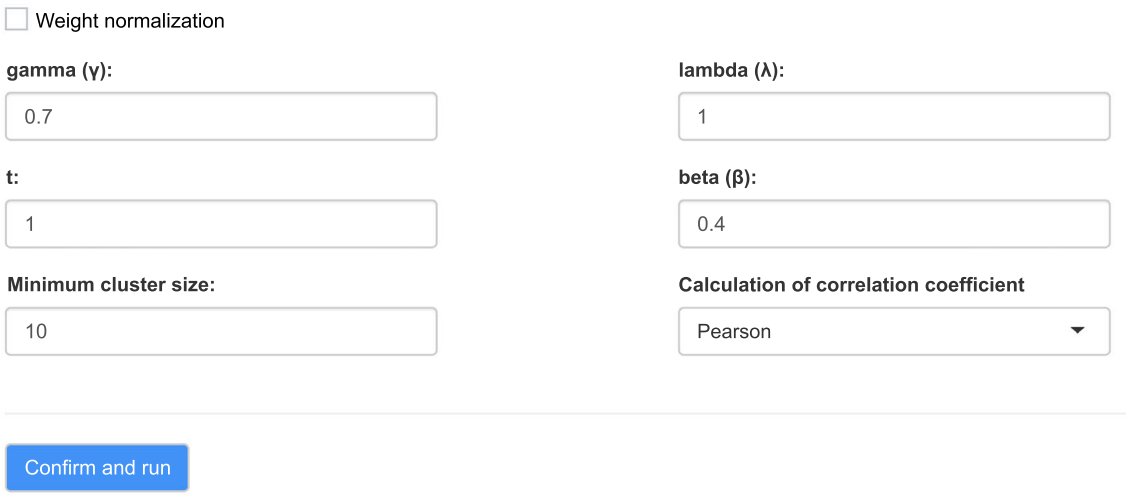


Figure 3 The lmQCM method panel
The lmQCM method panel that allows users to choose a variety of parameters. In this study, experiment with GSE17537 runs with unchecked weight normalization, $\gamma = 0.7$, $\lambda = 1$, $t = 1$, $\beta = 0.4$, minimum cluster size = 10, and Pearson correlation coefficient.

ensure all the values are numeric and can be processed by co-expression algorithms; 2) performing $\log_2(\chi+1)$ transformation of the expression values χ if the original values have not been previously transformed; 3) removing lowest J percentile rows (genes) with respect to mean expression values; and 4) removing lowest K percentile rows with respect to expression values' variance. These data filtering steps are necessary to reduce noise and to ensure the robustness for the downstream correlational computation in the lmQCM algorithm. The default settings are $J = 20$ and $K = 20$, by which genes with low expression and variance across samples are filtered out. In our example with GSE17537, we deselected logarithm conversion and NA value to 0 conversion and set $J = 50$ and $K = 10$, as shown in Figure 2B. However, users can always adjust these parameters based on their own needs and preferences. In the data pre-processing section, we further provide an "Advanced" panel to allow users to select a subgroup of samples of their interest. After the data pre-processing finishes, a dialog box appears to indicate how many genes are preserved after the filtering process.

Weighted network co-expression analysis

After data pre-processing, users can directly download pre-processed data or further proceed to the GCN analysis. In GCN analysis, we implemented the lmQCM algorithm as well as the WGCNA pipeline. The R package "WGCNA" from Bioconductor (http://bioconductor.org/) was adopted to integrate the WGCNA pipeline. We kept the mining steps concise and simple with default parameter settings, while preserving the flexibility for users to select parameters in each step. Guidelines for parameter selection are in the Method pages of the website. In addition, we also released the lmQCM package to CRAN (https://CRAN.R-project.org/package=lmQCM/).

In the lmQCM method panel, users can adjust parameters such as initial edge weight γ, weight threshold controlling parameters λ, t, and β, and the minimum cluster size (**Figure 3**). Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC) are implemented separately for users to select. SCC is recommended for analyzing RNA-seq data due to the large range of data values, and it is more robust than PCC to outliers. In our example with GSE17537, positive gene correlations were analyzed and the default settings were used (unchecked weight normalization, $\gamma = 0.7$, $\lambda = 1$, $t = 1$, $\beta = 0.4$, minimum cluster size = 10, and PCC for correlation measure). In the newer version of the TSUNAMI tool, both positive and negative correlations are considered during network module mining. The running time of lmQCM depends on the number of genes present after the filtering process. A progress bar is provided to show the program progress. Note that lmQCM will not work if the data contain no clustering structures or the gene pair correlations are so poor that none is above the initial mining starting threshold (γ). In those cases, the program will stop running and generate a warning message. However, this should not happen if the data contain enough highly correlated gene pairs after filtering and the default program settings are used.

The WGCNA method panel is a two-step analysis. Step 1 helps users to specify the hyper-parameter "power" in step 2, *i.e.*, the soft thresholding in Langfelder et al. [1] by visualizing the resulting plot (**Figure 4**A). Step 2 allows users to select the remaining parameters. TSUNAMI allows users to customize the parameters of power, reassign threshold, merge cut height, and indicate minimum module size. After applying WGCNA, a hierarchical clustering plot for the

resulting modules is also shown in this panel (Figure 4B). The resulting plot in Figure 4B is from the example data GSE17537 with powe = 10, reassign threshold = 0, merge cut height = 0.25, and minimum module size = 10.

In the last step of GCN mining, two outputs are provided by TSUNAMI: 1) merged gene clusters sorted by their sizes in descending order (**Figure 5**A with lmQCM algorithm) and 2) an eigengene matrix, which is the summarized expression values of genes in each GCN using the first principal component from singular value decomposition (Figure 5B with the lmQCM algorithm). Eigengene values can be regarded as the weighted average expression levels of each GCN. Such values are very useful for users to correlate GCN modules' expression profiles with various clinical and phenotypic traits in the downstream analysis, such as survival analysis. All results can be downloaded as files in CSV or TXT format.

Downstream enrichment analysis

Enrichr [23,24] is used as the tool for downstream gene set enrichment analysis implementation. By default, a total of 14 types of frequently used enrichment analyses are performed: 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; and 14) ChEA. Users can further customize the enrichment result categories from the open source code available in Github (https://github.com/huangzhii/TSUNAMI/).

To access Enrichr results, users can simply click the blue "GO" button in each row adjacent to the GCN mining results (as shown in Figure 5A). In each enrichment analysis, its output includes multiple results, such as the enriched term (*e.g.*, GO term or pathway), *P* value, Z-score, and overlapping genes. Users can download multiple analysis results that are bundled in a ZIP file. In addition, other popular gene set enrichment analysis websites are also directly linked in TSUNAMI to enhance convenience for users. In our example with GSE17537, we selected the 36th GCN module with 15 genes generated by lmQCM to be analyzed for enrichment, and each result table was sorted based on the *P* value generated by Enrichr. From the results in **Table 1**, we can see that the 36th GCN module is highly enriched in GO Biological Process term "type I interferon signaling pathway (GO:0060337)" (9 out of 148 genes).

Circos plot

TSUNAMI provides Circos plots [25] through intermediate results or inputs in the cases of human transcriptomic data. Circos plots are very useful graphs for visualizing the positions of genes on chromosomes and gene–gene relationships/interactions. The Circos plot function from the R package "circlize" [25] is adopted in this package for users to locate and visualize mined GCNs of human genes.

In TSUNAMI, users can visualize the Circos plot via the "Circos Plots" section, either by typing their own gene list separated by the carriage return character ("\n") directly, or by using the calculated GCN modules (*e.g.*, by clicking the yellow button right next to the "GO" button in Figure 5A). TSUNAMI supports both human genomes hg38 (GRCh38) and hg19 (GRCh37). To match the gene symbol to starting and ending sites on a chromosome, we use the *refGene* database downloaded from the UCSC genome browser [26]. If multiple starting/ending sites are matched, we choose the longest one with length calculated by:
length=|ending_site–staring_site|+1

By updating the plots, users can also choose the size of the plots and decide whether gene symbols and pair-wise

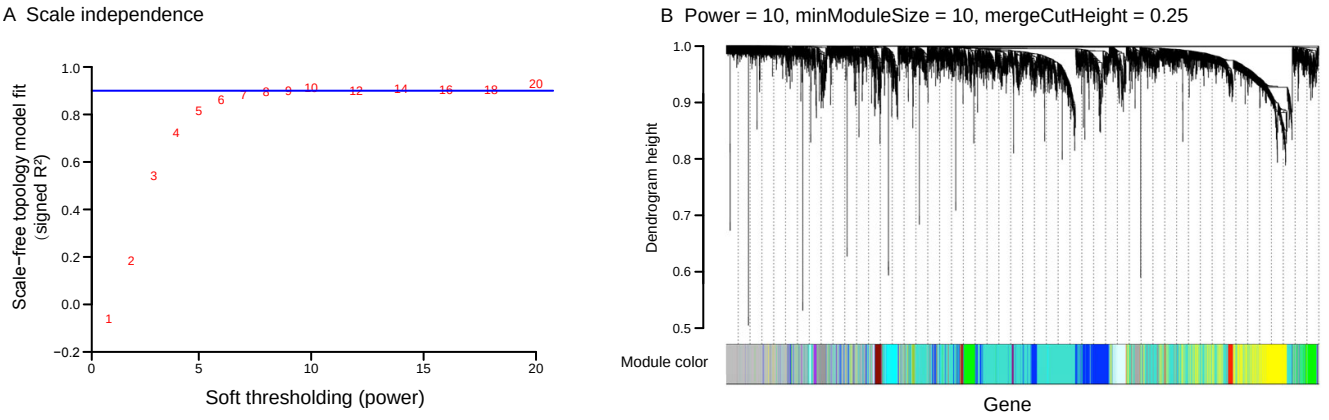


Figure 4 Choosing the power in WGCNA and the hierarchical clustering graph of WGCNA
A. The hyper-parameter "power" chosen from the value above the blue horizontal line. **B.** The hierarchical clustering graph with color bar indicating modules with GSE17537 dataset as an example. Parameters for WGCNA are power = 10, reassign threshold = 0, merge cut height = 0.25, and minimum module size = 10.

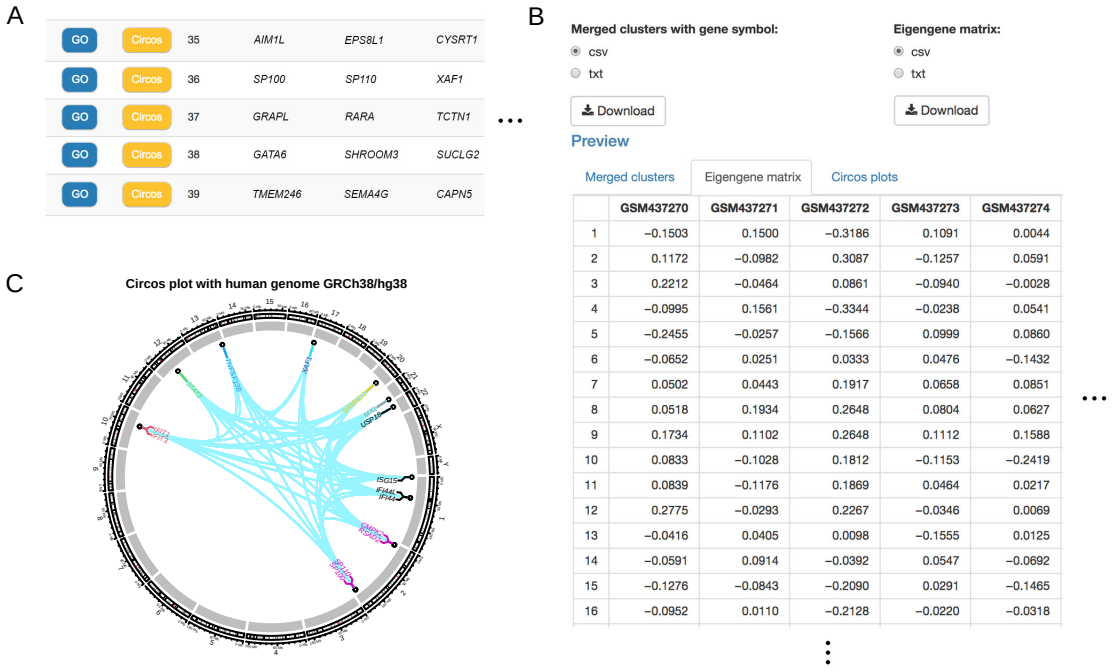


Figure 5 Merged cluster results generated by lmQCM

A. The merged GCN modules sorted in descending order based on the length of each cluster. The screenshot only shows part of the results (Clusters 35–39) with part of genes. **B.** The screenshot of the eigengene matrix (rounded to 4 decimal places for better visualization). Only part of the results (Clusters 1–16) with part of samples (GSM437270–GSM437274) are shown. **C.** The Circos plot results from the 36th GCN module with 15 genes. All modules in these subfigures are generated using the lmQCM algorithm with default parameters (unchecked weight normalization, $\gamma = 0.7$, $\lambda = 1$, $t = 1$, $\beta = 0.4$, minimum cluster size = 10, and Pearson correlation coefficient) with the GSE17537 dataset as an example.

Table 1 The partial results of GO enrichment analysis

ID	Term	Overlap	*P* value	Z-score	Overlapping gene
1	Type I interferon signaling pathway (GO:0060337)	9/148	2.51E−16	−3.2821	*SP100*; *RSAD2*; *STAT2*; *MX1*; *ISG15*; *SAMHD1*; *XAF1*; *IFIT1*; *IFIT3*
2	Cellular response to type I interferon (GO:0071357)	4/23	1.80E−09	−2.7766	*SP100*; *MX1*; *ISG15*; *IFIT1*
3	Negative regulation of single stranded viral RNA replication via double stranded DNA intermediate (GO:0045869)	4/44	2.73E−08	−2.6829	*RSAD2*; *MX1*; *ISG15*; *IFIT1*
4	Negative regulation of viral genome replication (GO:0045071)	4/40	1.84E−08	−2.4940	*RSAD2*; *MX1*; *ISG15*; *IFIT1*
5	Negative regulation by host of viral genome replication (GO:0044828)	4/51	5.01E−08	−2.6224	*RSAD2*; *MX1*; *ISG15*; *IFIT1*
6	Response to type I interferon (GO:0034340)	3/35	2.20E−06	−2.7155	*SP100*; *MX1*; *ISG15*
7	Regulation of type I interferon-mediated signaling pathway (GO:0060338)	3/43	4.14E−06	−2.6859	*STAT2*; *SAMHD1*; *USP18*
8	Negative regulation of type I interferon-mediated signaling pathway (GO:0060339)	2/43	4.66E−04	−2.5488	*STAT2*; *USP18*
9	Positive regulation of type I interferon-mediated signaling pathway (GO:0060340)	2/52	6.81E−04	−2.5122	*STAT2*; *USP18*
10	Positive regulation of Fas signaling pathway (GO:1902046)	1/7	5.24E−03	−2.9563	*SP100*

Note: This table contains partial rows and columns from original result (active panel: GO Biological Process) from the 36th GCN module with 15 genes generated by lmQCM with GSE17537 series matrix as data. GO terms are sorted by *P* value. We refer readers to explore other *P* values and scores from TSUNAMI webpage and Enrichr package. GO, Gene Ontology; GCN, gene co-expression network.

links should be shown on the graph.

An example output of the Circos plot is shown in Figure 5C. This example used the 36th GCN module with 15 genes from our previously discussed example (use a color set for texts to get a clear visual effect), annotated by gene symbols of human genome hg38 (GRCh38). The link between a pair of genes indicates that they belong to the same GCN module.

Circos plots can help users visualize the locations of genes in a GCN on human chromosomes, thus enabling

them to identify GCNs due to copy number variation and other structural changes. In the future, genome from mice and other species will be incorporated for Circos plots.

Survival analysis with respect to GCN modules

An optional step of survival analysis follows the generation of the eigengene matrix. It allows users to correlate the GCN modules' eigengenes with patient survival time (or event-free survival). This extension of the tool can be further customized to correlate module eigengenes with other clinical traits in the future version. In our current version, we only implements survival analysis as a starting point.

In the survival analysis, users can perform Overall Survival/Event-Free Survival (OS/EFS) analysis based on the GCN modules' eigengenes and look for GCN modules that are significantly associated with prognosis. Although, depending on the group of patients specified by users, such GCNs may not exist all the time. To carry out the analysis, users first select an eigengene (corresponding to a GCN module) in TSUNAMI. The program then splits the patients into two groups by the median of eigengenes. Next, it tests the two groups against OS/EFS by calculating the *P* value of the log-rank test [27,28]. Before doing so, users need to input the numerical survival time of OS/EFS (either in months or in days) with categorical events on OS/EFS status (1: deceased; 0: censored). The "*survdiff*" function from R package "*survival*" is adopted to calculate the *P* value and plot the Kaplan-Meier survival curves.

Taking GSE17537 with full survival information as an example, the Kaplan-Meier survival plot was generated according to the OS information by dichotomizing the 36th GCN module's eigengenes at its median to high and low groups, as shown in **Figure 6**. Such a GCN module was generated from the lmQCM method with default settings, as shown in Figure 3. This survival analysis offers researchers a tool to immediately identify any GCN modules that are associated with patients' survival time, thus allowing researchers to further study their roles as potential prognosis biomarkers, as well as the biological pathways that differentiate the patients.

Discussion

We released the online TSUNAMI tool package for gene co-expression module identification with direct link to the TCGA RNA-seq datasets and the NCBI GEO database, while also accommodating users' input data. It is a one-stop comprehensive tool package with several advantages, such as flexibility in parameter selections, comprehensive GCN mining tools, direct link to downstream gene set enrichment analysis, Circos plot visualization, and survival analysis, with downloadable results in each step. All of these features bring tremendous convenience to biological researchers.

In addition, TSUNAMI not only can process microarray, RNA-seq, and scRNA-seq transcriptomic data, but is also capable of processing any numerical valued matrix for weighted network module mining. If the users upload an adjacency matrix of any supported format with numerical values as the edge weights, TSUNAMI can be used to mine network modules. This extension will be implemented in version 2.0.

Code availability

The web service is freely accessible through URL: https://biolearns.medicine.iu.edu/. Source code is available at https://github.com/huangzhii/TSUNAMI/.

CRediT author statement

Zhi Huang: Methodology, Software, Formal analysis, Writing - original draft. **Zhi Han:** Methodology, Software, Formal analysis, Writing - original draft. **Tongxin Wang:** Resource, Data curation, Writing - review & editing. **Wei Shao:** Visualization, Data curation, Writing - review & editing. **Shunian Xiang:** Visualization, Data curation,

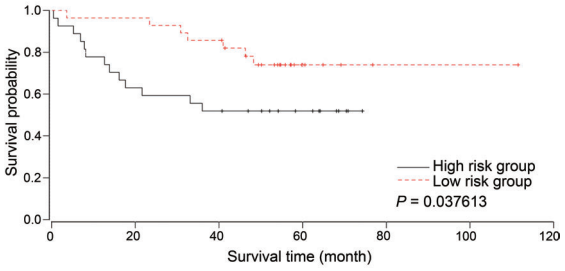


Figure 6 Survival analysis using GCN module eigengenes
Survival analysis using the 36th GCN module eigengenes generated from the lmQCM algorithm, with default parameters (unchecked weight normalization, $\gamma = 0.7$, $\lambda = 1$, $t = 1$, $\beta = 0.4$, minimum cluster size = 10, and Pearson correlation coefficient) with the GSE17537 dataset as an example. Fifty-five samples are used with overall survival information.

Writing - review & editing. **Paul Salama:** Project administration, Supervision, Writing - review & editing. **Maher Rizkalla:** Project administration, Supervision, Writing - review & editing. **Kun Huang:** Conceptualization, Project administration, Supervision, Funding acquisition, Writing - review & editing. **Jie Zhang:** Conceptualization, Project administration, Supervision, Methodology, Writing - original draft, Writing - review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work is partially supported by the American Cancer Society Inernal Reseatch Grant (to JZ), the National Cancer Institure Informatics Technology for Ccance Research U01 grant (Grant No. CA188547 to JZ and KH), and the Indiana University Precision Health Initiative (to JZ and KH). We thank the support from Indiana University Information Technologies and Advanced Biomedical IT Core. The results shown here are partially based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga/.

ORCID

0000-0001-6982-8285 (Zhi Huang)
0000-0002-5603-8433 (Zhi Han)
0000-0001-5826-1842 (Tongxin Wang)
0000-0003-1476-2068 (Wei Shao)
0000-0002-1351-0363 (Shunian Xiang)
0000-0002-7643-3879 (Paul Salama)
0000-0002-3723-8405 (Maher Rizkalla)
0000-0002-8530-370X (Kun Huang)
0000-0001-6939-7905 (Jie Zhang)

References

[1] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
[2] Zhang J, Huang K. Normalized ImQCM: an algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers. Cancer Inform 2016;13:137–46.
[3] Han Z, Johnson T, Zhang J, Zhang X, Huang K. Functional virtual flow cytometry: a visual analytic approach for characterizing single-cell gene expression patterns. Biomed Res Int 2017;2017:3035481.
[4] Han Z, Zhang J, Sun G, Liu G, Huang K. A matrix rank based concordance index for evaluating and detecting conditional specific co-expressed gene modules. BMC Genomics 2016;17:519.
[5] Zhang J, Huang K. Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. BMC Genomics 2017;18:1045.
[6] Chandran V, Coppola G, Nawabi H, Omura T, Versano R, Huebner EA, et al. A systems-level analysis of the peripheral nerve intrinsic axonal growth program. Neuron 2016;89:956–70.
[7] Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MPM, van Eijk K, et al. Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol 2012;13:R97.
[8] Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. Cancer Res 2017;77:e91–100.
[9] Shroff S, Zhang J, Huang K. Gene co-expression analysis predicts genetic variants associated with drug responsiveness in lung cancer. AMIA Jt Summits Transl Sci Proc 2016;2016:32–41.
[10] Zhang J, Abrams Z, Parvin JD, Huang K. Integrative analysis of somatic mutations and transcriptomic data to functionally stratify breast cancer patients. BMC Genomics 2016;17:513.
[11] Zhang J, Knobloch T, Parvin J, Weghorst C, Huang K. Identifying smoking associated gene co-expression networks related to oral cancer initiation. 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops 2011:1039–41.
[12] Zhang J, Lu K, Xiang Y, Islam M, Kotian S, Kais Z, et al. Weighted frequent gene co-expression network mining to identify genes involved in genome stability. PLoS Comput Biol 2012;8:e1002656.
[13] Zhang J, Ni S, Xiang Y, Parvin JD, Yang Y, Zhou Y, et al. Gene co-expression analysis predicts genetic aberration loci associated with colon cancer metastasis. Int J Comput Biol Drug Des 2013;6:60–71.
[14] Zhang J, Xiang Y, Ding L, Keen-Circle K, Borlawsky TB, Ozer HG, et al. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. BMC Bioinformatics 2010;11:S5.
[15] Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. Front Genet 2019;10:166.
[16] Xiang S, Huang Z, Wang T, Han Z, Yu CY, Ni D, et al. Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. BMC Med Genomics 2018;11:115.
[17] Yu CY, Xiang S, Huang Z, Johnson TS, Zhan X, Han Z, et al. Gene co-expression network and copy number variation analyses identify transcription factors involved in multiple myeloma progression. Front Genet 2019;10:468.
[18] Cao Y, Zhu J, Han G, Jia P, Zhao Z. scRNASeqDB: a database for gene expression profiling in human single cell by RNA-seq. bioRxiv 2017;104810.
[19] Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using *recount2*. Nat Biotechnol 2017;35:319–21.
[20] Freeman TJ, Smith JJ, Chen X, Washington MK, Roland JT, Means AL, et al. Smad4-mediated signaling inhibits intestinal neoplasia by inhibiting expression of beta-catenin. Gastroenterology 2012;142:562–71.
[21] Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 2010;138:958–68.
[22] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 2015;12:115–21.
[23] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 2013;14:128.
[24] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90–7.

[25] Gu Z, Gu L, Eils R, Schlesner M, Brors B. *circlize* implements and enhances circular visualization in R. Bioinformatics 2014;30:2811–2.

[26] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002;12:996–1006.

[27] Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ 1998;317:1572.

[28] Kleinbaum DG, Klein M. Kaplan-Meier survival curves and the log-rank test. In: Kleinbaum DG, Klein M, editors. Survival Analysis. Statistics for Biology and Health. New York: Spriger; 2012, p.55–96.