TSUNAMI: Translational Bioinformatics Tool Suite For Network Analysis And Mining

Zhi Huang^{1,2,4,#,a}, Zhi Han^{2,#,b}, Tongxin Wang^{3,c}, Wei Shao^{2,d}, Shunian Xiang^{5,6,e}, Paul Salama^{4,f}, Maher Rizkalla^{4,g}, Kun Huang^{2,*,h}, Jie Zhang^{6,*,i}

¹ School of Electrical and Computer Engineering, Purdue University, West Lafayette IN 47907, USA

² Department of Medicine, Indiana University School of Medicine, Indianapolis IN 46202, USA

³ Department of Computer Science, Indiana University, Bloomington IN 47405, USA

⁴ Department of Electrical and Computer Engineering, Indiana University - Purdue University Indianapolis, Indianapolis IN 46202, USA

⁵ School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China

⁶ Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis IN 46202, USA

Equal contribution.

* Corresponding authors.

E-mail: jjzhan@iu.edu (Zhang J), kunhuang@iu.edu (Huang K).

Running title: Huang Z *et al* / A tool package for correlational network module mining and functional analysis

^aORCID: 0000-0001-6982-8285.

^bORCID: 0000-0002-5603-8433.

^cORCID: 0000-0001-5826-1842.

^dORCID: 0000-0003-1476-2068.

^eORCID: 0000-0002-1351-0363.

^fORCID: 0000-0002-7643-3879.

^gORCID: 0000-0002-3723-8405.

^hORCID: 0000-0002-8530-370X.

ⁱORCID: 0000-0001-6939-7905.

Total word counts (from “Introduction” to “Conclusion”): 2613

Total figures: 6

Total tables: 1

Total supplementary figures: 0

Total supplementary tables: 0

Total supplementary files: 0

Abstract

Gene co-expression network (GCN) mining identifies gene modules with highly correlated expression profiles across samples/conditions. It helps to discover latent gene/molecular interactions, identify novel gene functions, and extract molecular features from certain disease/condition groups, thus help 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 Ontology (GO) enrichment analysis, as well as modules that may share common members. To address this need, we develop a GCN mining tool package TSUNAMI (Tools SUite for Network Analysis and MIning) which incorporates our state-of-the-art ImQCM algorithm to mine GCN modules in public and user-input data (microarray, RNA-seq, or any other numerical omics data), then performs downstream GO and enrichment analysis based on the modules identified. It has several features and advantages: (i) user friendly interface and the real-time co-expression network mining through web server; (ii) direct access and search of GEO and TCGA databases as well as user-input expression matrix (microarray, RNA-seq, etc.) for GCN module mining; (iii) multiple co-expression analysis tools to choose with highly flexible of parameter selection options; (iv) identified GCN modules are summarized to eigengenes, which are convenient for user to check their correlation with other clinical traits; (v) integrated downstream Enrichr enrichment analysis and links to other GO tools; (vi) visualization of gene loci by Circos plot in any step. The web service is freely accessible through URL: <http://spore.ph.iu.edu:3838/zhihuan/TSUNAMI/>. Source code is available at <https://github.com/huangzhii/TSUNAMI/>.

KEYWORDS: Network mining; Gene co-expression network; Transcriptomic data analysis; ImQCM; Web server

Introduction

Gene co-expression network (GCN) mining is a popular bioinformatics approach to identify densely connected gene modules, which are linked by their highly correlated expression profiles. It helps reveal latent gene/molecule interactions, identify novel gene functions, disease pathways and biomarkers, as well as provide disease mechanistic insights. GCN mining approaches such as WGCNA [1] and lmQCM [2] have been used increasingly [3–7]. Compared to the more popularly used WGCNA package, lmQCM is capable to mine smaller densely co-expressed GCN modules and allow overlapped membership in the output modules. Those features reflect closely the real biological networks *in vivo*, where the same genes may participate in multiple pathways and a small group of genes are more likely to be synergistically regulated in local pathway functions. Besides, the generally smaller size of modules from lmQCM usually generates more meaningful GO enrichment results, which has been successfully applied to many diseases and cancer types [8–17].

There exist several online databases that curate transcriptomic data, for example, PanglaoDB (<https://panglaodb.se/>) collected single-cell RNA-seq (scRNA-seq) data from mouse and human. Cao et al. scRNASeqDB [18] provides an scRNA-seq database for gene expression profiling in human. Recount2 [19] provides public available analysis-ready gene and exon counts datasets. However, all of these databases focus on data collection, to the best of our knowledge, there is no tool offering the entire pipeline that can directly process transcriptomic data, mine GCN modules, analyse GO enrichment, and visualized the results in a complete pipeline fashion. To meet such needs, we released our web-based analysis tool suite TSUNAMI (Tools SUite for Network Analysis and MIning).

For users' convenience, TCGA mRNA-seq data (Illumina HiSeq RSEM genes normalized from <https://gdac.broadinstitute.org/>) and NCBI Gene Expression Omnibus (GEO) are directly incorporated into TSUNAMI, where GEO contains a large number of microarray datasets. In addition, other data types such as miRNA-seq and DNA methylation are also compatible with this suite. In fact, TSUNAMI can handle any numerical matrix data regardless which omics data type it is from. In TSUNAMI, it not only incorporates the newly released lmQCM algorithm, but also includes WGCNA package for users to explore and compare their GCN modules from

two different algorithms. We offer highly flexible parameter choices in each step to users who may want to fine tune each algorithm to suit for their own data and goal.

Prior to data mining, a data pre-processing interface is designed to address the input data format difference and filter the data to remove noise for GCN mining. Each step of the pre-processing is transparent to users and can be adjusted according to their own preferences and needs.

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

Data input

A flowchart that describes 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 with Python to improve the computing speed in module mining step. Some front-end interfaces and functions are done by JavaScript. In TSUNAMI, users can choose to use either TCGA RNA-seq expression data, GSE series matrix data, or other RNA-seq data from GEO database, or local user-input numerical matrix data, such as microarray, RNA-seq, scRNA-seq data, DNA methylation data, or any other type of numeric matrix data. User can also choose specific omics data type on GEO database if keywords are given to indicate the data type in the search window. Only few GSE data is not able to be processed (for example, 12 out of first 1000 GSE data), mostly are legacy microarray data, which contain too much missing data or too small sample size. Other 98.80% of first 1000 GSE data can be processed. On the website, various of example data from microarray to scRNA-seq data are listed on TSUNAMI for users' reference. Instead of searching GEO database manually, TSUNAMI provides a friendly interface for users to retrieve data from GEO by keywords, offers flexible select tool to retrieve relevant GSE dataset to perform GCN analysis.

TSUNAMI also provides an upload bar for users to upload local files in various formats (CSV, TSV, XLSX, TXT, etc.), the upload interface is shown in **Figure 2A**. In this paper, one microarray dataset (GSE17537 from GEO) is chosen as an example to present the features of TSUNAMI. GSE17537 contains microarray data of 55 colorectal cancer patients from Vanderbilt Medical Center (VMC), with 54,675 probesets [20, 21].

Online data pre-processing

One issue of GEO microarray data is that different platforms adopted 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 GSE17537, the annotation platform is GPL570. TSUNAMI can also automatically identify annotation platforms of the data from GEO. During the conversion, TSUNAMI will (i) remove rows with empty gene symbol; and (ii) select the rows with the largest mean expression value when multiple probesets are matched with the same gene symbol. The interface of data pre-processing step is shown in **Figure 2B**.

Additional data filtering steps include: (i) convert “NA” value (not a number value) to 0 in expression data, to ensure all the values are numeric and can be interpreted by co-expression algorithms; (ii) perform $\log_2(x + 1)$ transformation of the expression values x if the original values have not been transformed previously; (iii) remove lowest J percentile rows (genes) with respect to mean expression values; (iv) remove 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 lmQCM algorithm. The default settings are $J = 50$ and $K = 50$, by which genes with low expression and variance across samples are filtered out. In our example with GSE17537, we deselect logarithm conversion and NA value to 0 conversion, 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 Data Pre-processing section, we further provide “Advanced” panel to allow users select samples subgroup of their interest. After finished the data pre-processing, a dialog box will appear to indicate how many genes left after the filtering process.

99

100 **Weighted network co-expression analysis**

101 After data pre-processing, users can directly download pre-processed data or further
 102 proceed to GCN analysis. In GCN analysis, we implemented lmQCM algorithm as
 103 well as WGCNA pipeline. We kept the mining steps concise and simple with default
 104 parameter settings, while preserving the flexibility for users to select parameters in
 105 each step. Guidelines for parameter selection are in method pages of the website.
 106 Besides this article, we also release the lmQCM package to CRAN
 107 (<https://CRAN.R-project.org/package=lmQCM>). The R package “WGCNA” from
 108 Bioconductor (<http://bioconductor.org>) was adopted to integrate the WGCNA
 109 pipeline.

110 In the lmQCM method panel, users can adjust parameters such as initial edge
 111 weight γ , weight threshold controlling parameters λ , t , β , and minimum cluster
 112 size (**Figure 3**). Pearson correlation coefficient (PCC) and Spearman's rank
 113 correlation coefficient (SCC) are implemented separately for users to select. SCC is
 114 recommended for analysing RNA-seq data due to the large range of data values, and it
 115 is more robust than PCC to tolerate outliers. In our example with GSE17537, the
 116 default settings were used (unchecked weight normalization, $\gamma = 0.7$, $\lambda = 1$, $t = 1$,
 117 $\beta = 0.4$, minimum cluster size= 10, and PCC for correlation measure). The running
 118 time of lmQCM depends on the number of genes after filtering process. A progress
 119 bar is provided to show the program progress. Note that lmQCM will not work if the
 120 data contain no clustering structure or the gene pair correlations are so poor that none
 121 is above the initial mining starting threshold (γ). In those cases, the program will stop
 122 running and generate a warning message. However, if the data contain enough high
 123 correlated gene pairs after filtering and with the default program settings, this should
 124 not happen.

125 The WGCNA method panel is a two-step analysis: Step 1 helps users to specify
 126 the hyper-parameter “power” in step 2, *i.e.*, the soft thresholding in [1] by visualizing
 127 the resulting plot (**Figure 4A**). Step 2 allows users to select the remaining parameters.
 128 TSUNAMI allows users to customize the parameters of power, reassign threshold,
 129 merge cut height, and minimum module size. After applying WGCNA, a hierarchical
 130 clustering plot for getting the result modules is also shown in this panel (**Figure 4B**).
 131 The resulting plot in **Figure 4B** is from the example data GSE17537 with power= 10,

132 set reassign threshold = 0, merge cut height = 0.25, and minimum module size
133 = 10.

134 In the last step of GCN mining, two outputs are provided by TSUNAMI: (i)
135 merged gene clusters sorted by their sizes in descending order (**Figure 5A** with
136 lmQCM algorithm); (ii) an eigengene matrix, which is the expression values of each
137 GCN summarized into the first principal component using singular value
138 decomposition (**Figure 5C** with lmQCM algorithm). Eigengene values can be
139 regarded as the weighted average expressions of each GCN, thus each GCN is
140 summarized to a “super gene” with the first right singular vector as the expression
141 values. Such values are very useful for users to correlate GCN modules expression
142 profiles with various traits in the downstream analysis such as survival analysis. All
143 results can be downloaded in CSV or TXT format.

144

145 **Downstream enrichment analysis**

146 Enrichr [23, 24] is used as the tool for downstream GO enrichment analysis
147 implementation. By default, total 14 types of frequent used enrichment are performed.
148 They are (1) Biological Process; (2) Molecular Function; (3) Cellular Component; (4)
149 Jensen DISEASES; (5) Reactome; (6) KEGG; (7) Transcription Factor PPIs; (8)
150 Genome Browser PWMs; (9) TRANSFAC and JASPAR PWMs; (10) ENCODE TF
151 ChIP-seq; (11) Chromosome Location (Cytoband); (12) miRTarBase; (13)
152 TargetScan microRNA; (14) ChEA. Users can further customize the enrichment result
153 categories from the open source code available in Github
154 (<https://github.com/huangzhii/TSUNAMI>).

155 To access Enrichr results, users can simply click the blue button “GO” in each
156 row adjacent to the GCN mining results (as shown in **Figure 5A**). In each enrichment
157 analysis result, it outputs the term (e.g., GO or pathway), *P* value, z-score, overlapped
158 genes, etc. Users can download multiple analysis results which are bundled in a ZIP
159 file. Besides, other popular GO analysis websites are also directly linked in
160 TSUNAMI to bring conveniences to users. In our example with GSE17537, we select
161 the 36th GCN module with 15 genes generated by lmQCM to analyze the GO
162 enrichment, and each result table are sorted based on the *P* value that Enrichr
163 calculated. From the result in **Table 1**, we can see the 36th GCN module is highly
164 overlapped with GO Biological Process term “type I interferon signaling pathway
165 (GO:0060337)” (9 out of 148 genes).

166

167 **Circos plot**

168 TSUNAMI provides Circos plots [25] through any intermediate results or inputs in
169 the cases of human transcriptomic data. Circos plot is a very useful graph for
170 visualizing the positions of genes on chromosomes and gene-gene
171 relationships/interactions. The Circos plot function from the R package “circlize” [25]
172 is adopted in this package for users to locate and visualize mined GCNs of human
173 genes.

174 In TSUNAMI, users can visualize the Circos plot via “Circos Plots” section, either
175 by typing their own genes list separated by carriage return character (“\n”) directly, or
176 using the calculated GCN modules (for example, by clicking the yellow button right
177 next to the “GO” button in **Figure 5A**). TSUNAMI supports both human genomes
178 hg38 (GRCh38) and hg19 (GRCh37). To match the gene symbol to chromosomes’
179 starting and ending sites, we use reference gene table from UCSC genome browser
180 [26]. If multiple starting/ending site are matched, we choose the longest one with
181 length calculated by:

$$182 \text{ length} = \text{ending_site} - \text{starting_site} + 1 \quad (1)$$

183 By updating the plots, users can also choose the size of the plots and decide
184 whether gene symbols and pair-wised links should be shown on the graph.

185 An example output of Circos plot in **Figure 5B** used the 36th GCN module with
186 15 genes in the lmQCM result from GSE17537 series matrix (use a color set for texts
187 to get a clear visual effect), indicated by gene symbols of human genome hg38
188 (GRCh38). While the link between a pair of genes indicates that they belong to the
189 same co-expressed GCN module.

190 Circos plots can help users to visualize the GCN module’s location on human
191 chromosomes from either lmQCM or WGCNA mining, help them to visualize GCNs
192 due to copy number variation and other structural changes. In the future, genome from
193 mouse and other species will be incorporated for Circos plot.

194

195 **Survival analysis with respect to GCN modules**

196 An optional step of survival analysis follows the generation of the eigengene matrix.
197 It allows users to correlate the GCN module’s eigengene values with patient clinical
198 survival (or event-free survival), and such extension tool can be further customized as

users' need to correlate module eigengene values with other clinical traits in the future version. In our current version, we only implemented survival analysis as an example.

In the survival analysis, users can perform Overall Survival/Event-Free Survival (OS/EFS) analysis based on the GCN modules' eigengene values, and look for significant GCNs that are capable for prognosis, although depending on the group of patients user specifies, such GCNs may not be identified all the time. TSUNAMI lets user to select an eigengene row (corresponding to a GCN module). The program will splits the patients into two groups by eigengene values' median, then tests 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 OS/EFS status (1: deceased; 0: censored). "survdif" function from R package "survival" is adopted to calculate the P value and plot the Kaplan-Meier survival curve.

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

Conclusion

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

Besides, TSUNAMI can not only process microarray, RNA-seq, and single-cell RNA-seq transcriptomic data, but also be capable for processing any type of the numerical valued matrix for weighted network module mining. If the users upload an

232 adjacency matrix of any supported format with numerical values as the edge weights,
233 TSUNAMI can be used to mine any correlational network modules or even beyond
234 that. This extension will be implemented in version 2.0.

235

236 **Authors' contributions**

237 JZ and KH conceived the idea of the project and participated in software design and
238 helped to draft the manuscript. Zhi Huang and Zhi Han wrote the software and
239 manuscript. TW, WS, and SX carried out the GO enrichment analysis tool options.
240 PS, MR, KH, JZ provide research guidance. JZ and KH reviewed and edited the
241 manuscript. All authors read and approved the final manuscript.

242

243 **Competing interests**

244 The authors have declared no competing interests.

245

246 **Acknowledgements**

247 This work is partially supported by IUSM startup fund (JZ), ACS IRG grant (JZ), the
248 NCI ITCR U01 (CA188547) grant (JZ and KH), and Indiana University Precision
249 Health Initiative (JZ and KH). We thank the support from Indiana University
250 Information Technologies and Advanced Biomedical IT Core. The results shown here
251 are partially based upon data generated by the TCGA Research Network:
252 <https://www.cancer.gov/tcga>.

253 **References**

- 254 [1] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation
255 network analysis. *Bmc Bioinformatics* 2008;9.
- 256 [2] Zhang J, Huang K. Normalized ImQCM: An Algorithm for Detecting Weak
257 Quasi-Cliques in Weighted Graph with Applications in Gene Co-Expression Module
258 Discovery in Cancers. *Cancer informatics* 2014;13:CIN. S14021.
- 259 [3] Han Z, Johnson T, Zhang J, Zhang X, Huang K. Functional Virtual Flow
260 Cytometry: A Visual Analytic Approach for Characterizing Single-Cell Gene
261 Expression Patterns (vol 2017, 3035481, 2017). *Biomed Research International* 2017.
- 262 [4] Han Z, Zhang J, Sun GY, Liu G, Huang K. A matrix rank based concordance index
263 for evaluating and detecting conditional specific co-expressed gene modules. *Bmc*
264 *Genomics* 2016;17.
- 265 [5] Zhang J, Huang K. Pan-cancer analysis of frequent DNA co-methylation patterns
266 reveals consistent epigenetic landscape changes in multiple cancers. *Bmc Genomics*
267 2017;18.
- 268 [6] Chandran V, Coppola G, Nawabi H, Omura T, Versano R, Huebner EA, et al. A
269 Systems-Level Analysis of the Peripheral Nerve Intrinsic Axonal Growth Program.
270 *Neuron* 2016;89:956-70.
- 271 [7] Horvath S, Zhang YF, Langfelder P, Kahn RS, Boks MPM, van Eijk K, et al.
272 Aging effects on DNA methylation modules in human brain and blood tissue.
273 *Genome Biology* 2012;13.
- 274 [8] Cheng J, Zhang J, Han YT, Wang XS, Ye XF, Meng YB, et al. Integrative Analysis
275 of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell
276 Carcinoma Prognosis. *Cancer Research* 2017;77:E91-E100.
- 277 [9] Shroff S, Zhang J, Huang K. Gene Co-Expression Analysis Predicts Genetic
278 Variants Associated with Drug Responsiveness in Lung Cancer. *AMIA Jt Summits*
279 *Transl Sci Proc* 2016;2016:32-41.
- 280 [10] Zhang J, Abrams Z, Parvin JD, Huang K. Integrative analysis of somatic
281 mutations and transcriptomic data to functionally stratify breast cancer patients. *Bmc*
282 *Genomics* 2016;17.
- 283 [11] Zhang J, Knobloch T, Parvin J, Weghorst C, Huang K. Identifying Smoking
284 Associated Gene Co-expression Networks Related to Oral Cancer Initiation. 2011
285 *Ieee International Conference on Bioinformatics and Biomedicine Workshops*
286 2011:1039-41.
- 287 [12] Zhang J, Lu KW, Xiang Y, Islam M, Kotian S, Kais Z, et al. Weighted Frequent
288 Gene Co-expression Network Mining to Identify Genes Involved in Genome Stability.
289 *Plos Computational Biology* 2012;8.
- 290 [13] Zhang J, Ni S, Xiang Y, Parvin JD, Yang Y, Zhou Y, et al. Gene Co-expression
291 analysis predicts genetic aberration loci associated with colon cancer
292 metastasis 2013;6:60-71.
- 293 [14] Zhang J, Xiang Y, Ding L, Borlawsky TB, Ozer HG, Jin R, et al. Using gene
294 co-expression network analysis to predict biomarkers for chronic lymphocytic
295 leukemia. *BMC bioinformatics* 2010;11:S5.
- 296 [15] Huang Z, Zhan X, Xiang S, Johnson T, Helm B, Yu C, et al. SALMON: Survival
297 Analysis Learning with Multi-Omics Neural Networks on Breast Cancer. *Frontiers in*
298 *Genetics* 2019;10:166.
- 299 [16] Xiang S, Huang Z, Wang T, Han Z, Yu CY, Ni D, et al. Condition-specific gene
300 co-expression network mining identifies key pathways and regulators in the brain
301 tissue of Alzheimer's disease patients. *BMC Med Genomics* 2018;11:115.

302 [17] Yu CY, Xiang S, Huang Z, Johnson TS, Zhan X, Han Z, et al. Gene
303 Co-expression Network and Copy Number Variation Analyses Identify Transcription
304 Factors Involved in Multiple Myeloma Progression. *Frontiers in genetics*
305 2019;10:468.

306 [18] Cao Y, Zhu J, Han G, Jia P, Zhao Z. scRNASeqDB: a database for gene
307 expression profiling in human single cell by RNA-seq. *bioRxiv* 2017:104810.

308 [19] Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al.
309 Reproducible RNA-seq analysis using recount2. *Nature Biotechnology*
310 2017;35:319-21.

311 [20] Freeman TJ, Smith JJ, Chen X, Washington MK, Roland JT, Means AL, et al.
312 Smad4-Mediated Signaling Inhibits Intestinal Neoplasia by Inhibiting Expression of
313 beta-Catenin. *Gastroenterology* 2012;142:562-U228.

314 [21] Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang AX, et al.
315 Experimentally Derived Metastasis Gene Expression Profile Predicts Recurrence and
316 Death in Patients With Colon Cancer. *Gastroenterology* 2010;138:958-68.

317 [22] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al.
318 Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*
319 2015;12:115-21.

320 [23] Chen EY, Tan CM, Kou Y, Duan QN, Wang ZC, Meirelles GV, et al. Enrichr:
321 interactive and collaborative HTML5 gene list enrichment analysis tool. *Bmc*
322 *Bioinformatics* 2013;14.

323 [24] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan QN, Wang ZC, et
324 al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.
325 *Nucleic Acids Research* 2016;44:W90-W7.

326 [25] Gu ZG, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances
327 circular visualization in R. *Bioinformatics* 2014;30:2811-2.

328 [26] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The
329 human genome browser at UCSC. *Genome Research* 2002;12:996-1006.

330 [27] Bland JM, Altman DG. Statistics notes - Survival probabilities (the Kaplan-Meier
331 method). *British Medical Journal* 1998;317:1572-.

332 [28] Kleinbaum DG, Klein M. Kaplan-Meier Survival Curves and the Log-Rank Test.
333 *Survival Analysis: A Self-Learning Text, Third Edition* 2012:55-96.

334

335 **Figure legends**

336 **Figure 1 Flowchart of TSUNAMI.**

337 In this flowchart representation of TSUNAMI pipeline, blue rectangles represent
338 pipeline operations; rounded rectangles in pink represent download processes.

339 **Figure 2 Dataset Selection and Pre-processing Panel**

340 **A.** Data can be uploaded manually, or chosen from NCBI GEO database (not shown
341 in the figure). When uploading the data, the maximum file size that TSUNAMI allows
342 is 300 Megabytes. Header, separators and quote methods can be adjusted by users. **B.**
343 The Data Pre-processing Panel includes several pre-processing steps.

344 **Figure 3 lmQCM Method Panel Data Pre-processing Panel.**

345 The lmQCM algorithm panel which allows users to choose various of parameters. In
346 this paper, experiment runs with unchecked weight normalization, $\gamma = 0.7$, $\lambda = 1$,
347 $t = 1$, $\beta = 0.4$, minimum cluster size = 10, and adopted Pearson correlation
348 coefficient.

349 **Figure 4 Choosing the Power in WGCNA and the Hierarchical Clustering** 350 **Graph of WGCNA**

351 **A.** The hyper-parameter “power” that chosen from the value above the blue horizontal
352 line. **B.** The result hierarchical clustering graph with color bar indicating result
353 modules with GSE17537 series matrix as an example, use parameters power= 10,
354 reassign threshold = 0, merge cut height = 0.25, minimum module size = 10 in
355 WGCNA.

356 **Figure 5 Merged Clusters Result Generated by lmQCM**

357 **A.** The merged GCN module results, sorted in descending order based on the length
358 of each cluster. Figure only shows part of the results (cluster 35~39) with part of
359 genes. **B.** The Circos plot result from the 36th GCN module with 15 genes. **C.** The
360 screenshot of the eigengene matrix (rounded to 4 decimal places for better
361 visualization). Figure only shows part of the results (cluster 1~16) with part of
362 samples (GSM437270~GSM437274). All subfigures use lmQCM algorithm with
363 default parameters (unchecked weight normalization, $\gamma = 0.7$, $\lambda = 1$, $t = 1$,
364 $\beta = 0.4$, minimum cluster size = 10, and adopted Pearson correlation coefficient)
365 with GSE17537 series matrix as an example.

366 **Figure 6 Survival Analysis using GCN Module Eigenvalues**

367 Survival analysis using the 36th GCN module eigenvalues generated from lmQCM
 368 algorithm, with default parameters (unchecked weight normalization, $\gamma = 0.7$, $\lambda =$
 369 1, $t = 1$, $\beta = 0.4$, minimum cluster size = 10, and adopted Pearson correlation
 370 coefficient) with GSE17537 series matrix as an example. 55 samples are used with
 371 Overall Survival information.

372 **Tables**

373 **Table 1 The partial results of GO enrichment analysis**

374 *Note:* This table contains partial rows and columns from original result (active panel:
 375 GO Biological Process) from the 36th GCN module with 15 genes generated by
 376 lmQCM with GSE17537 series matrix as data. GO terms are sorted by *P* value. We
 377 refer readers to explore other *P* values and scores from TSUNAMI webpage and
 378 Enrichr package.

A

File Uploader

Choose File

Browse...

No file selected

Note: Maximum file size allowed for uploading is 300MB. If uploaded data is with .xlsx or .xls, separator can be any value, but please make sure data are located in Sheet1.

☒ Header

Separator

☒ Comma

☐ Semicolon

☐ Tab

☐ Space

Quote

☐ None

☒ Double Quote

☐ Single Quote

B

Basic

Advanced

Verify starting column and row of expression data

Choose starting column and row for expression data.

Default value when leave them blank: starting row = 1, starting column = 2.

Gene and Expression starting row:

Expression starting column:

1

2

Convert Probe ID to Gene Symbol

Convert Probe ID to Gene Symbol with Platform GPL*** (Optional for self-uploaded data):

Be sure to verify (modify) Gene Symbol.

GPL570

Remove Genes

Remove data with lowest percentile mean expression value shared by all samples. Then remove data with lowest percentile variance across samples.

Default value when leave them blank: 0.

Lowest Mean Percentile (%) To Remove:

50

Lowest Variance Percentile (%) To Remove:

10

☐ Convert NA value to 0 in Expression Data

☐ Take the $\log_2(x+1)$ of Expression Data x (Default: Unchecked)

☒ Remove rows with empty Gene Symbol

☒ Keep only one row with largest mean expression value when Gene Symbol is duplicated

☐ Weight Normalization

gamma (γ):

0.7

t:

1

Minimum Cluster Size:

10

lambda (λ)

1

beta (β):

0.4

Calculation of Correlation Coefficient

pearson

Confirm and Run

A

B

A

GO	Circos	35	AIM1L	EP58L1	CYSRT1
GO	Circos	36	SP100	SP110	XAF1
GO	Circos	37	GRAPL	RARA	TCTN1
GO	Circos	38	GATA8	SHROOM3	SUCLG2
GO	Circos	39	TMEM246	SEMA4G	CAPN5

...

B

C

Merged Clusters with Gene Symbol:

- ☒ csv
- ☐ txt

Download

Eigengene Matrix:

- ☒ csv
- ☐ txt

Download

Preview

Merged Clusters

Eigengene Matrix

Circos Plots

	GSM437270	GSM437271	GSM437272	GSM437273	GSM437274
1	-0.1503	0.1500	-0.3186	0.1091	0.0044
2	0.1172	-0.0982	0.3087	-0.1257	0.0591
3	0.2212	-0.0464	0.0861	-0.0940	-0.0028
4	-0.0995	0.1561	-0.3344	-0.0238	0.0541
5	-0.2455	-0.0257	-0.1588	0.0999	0.0860
6	-0.0652	0.0251	0.0333	0.0476	-0.1432
7	0.0502	0.0443	0.1917	0.0658	0.0851
8	0.0518	0.1934	0.2648	0.0804	0.0627
9	0.1734	0.1102	0.2648	0.1112	0.1588
10	0.0833	-0.1028	0.1812	-0.1153	-0.2419
11	0.0839	-0.1176	0.1869	0.0464	0.0217
12	0.2775	-0.0293	0.2267	-0.0346	0.0069
13	-0.0416	0.0405	0.0098	-0.1555	0.0125
14	-0.0591	0.0914	-0.0392	0.0547	-0.0692
15	-0.1278	-0.0843	-0.2090	0.0291	-0.1465
16	-0.0952	0.0110	-0.2128	-0.0220	-0.0318

...

...

Survival analysis using GCN module eigenvalues

Black line: high risk group; Red dashed line: low risk group; p-value: 0.037613